



HAESE MATHEMATICS

Mathematics

for the international student

Mathematics III (Option):
Probability

LANZAROTE
Caliente.COM

Topic 7

FM Topic 3



Catherine Quinn
Peter Blythe
Robert Haese
Michael Haese

for use with

IB Diploma Programme

Mathematics

for the international student
Mathematics HL (Option):
Statistics and Probability

HL Topic 7

FM Topic 3



Catherine Quinn

Peter Blythe

Robert Haese

Michael Haese

for use with
IB Diploma
Programme

MATHEMATICS FOR THE INTERNATIONAL STUDENT

Mathematics HL (Option): Statistics and Probability

Catherine Quinn	B.Sc.(Hons), Grad.Dip.Ed., Ph.D.
Peter Blythe	B.Sc.
Robert Haese	B.Sc.
Michael Haese	B.Sc.(Hons.), Ph.D.

Haese Mathematics
152 Richmond Road, Marlestone, SA 5033, AUSTRALIA
Telephone: +61 8 8210 4666, Fax: +61 8 8354 1238
Email: info@haesemathematics.com.au
Web: www.haesemathematics.com.au

National Library of Australia Card Number & ISBN 978-1-921972-31-7

© Haese & Harris Publications 2013

Published by Haese Mathematics.
152 Richmond Road, Marlestone, SA 5033, AUSTRALIA

First Edition 2013

Artwork by Brian Houston.

Cover design by Piotr Poturaj.

Typeset in Australia by Deanne Gallasch. Typeset in Times Roman 10 $\frac{1}{2}$.

Printed in Malaysia through Bookpac Production Services, Singapore.

The textbook and its accompanying CD have been developed independently of the International Baccalaureate Organization (IBO). The textbook and CD are in no way connected with, or endorsed by, the IBO.

This book is copyright. Except as permitted by the Copyright Act (any fair dealing for the purposes of private study, research, criticism or review), no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher. Enquiries to be made to Haese Mathematics.

Copying for educational purposes: Where copies of part or the whole of the book are made under Part VB of the Copyright Act, the law requires that the educational institution or the body that administers it has given a remuneration notice to Copyright Agency Limited (CAL). For information, contact the Copyright Agency Limited.

Acknowledgements: While every attempt has been made to trace and acknowledge copyright, the authors and publishers apologise for any accidental infringement where copyright has proved untraceable. They would be pleased to come to a suitable agreement with the rightful owner.

Disclaimer: All the internet addresses (URLs) given in this book were valid at the time of printing. While the authors and publisher regret any inconvenience that changes of address may cause readers, no responsibility for any such changes can be accepted by either the authors or the publisher.

FOREWORD

Mathematics HL (Option): Statistics and Probability has been written as a companion book to the Mathematics HL (Core) textbook. Together, they aim to provide students and teachers with appropriate coverage of the two-year Mathematics HL Course, to be first examined in 2014.

This book covers all sub-topics set out in Mathematics HL Option Topic 7 and Further Mathematics HL Topic 3, Statistics and Probability.

The aim of this topic is to introduce students to the basic concepts and techniques of statistics and probability and their applications.

Detailed explanations and key facts are highlighted throughout the text. Each sub-topic contains numerous Worked Examples, highlighting each step necessary to reach the answer for that example.

Theory of Knowledge is a core requirement in the International Baccalaureate Diploma Programme, whereby students are encouraged to think critically and challenge the assumptions of knowledge. Discussion topics for Theory of Knowledge have been included on pages 157 to 159. These aim to help students discover and express their views on knowledge issues.

The accompanying student CD includes a PDF of the full text and access to specially designed graphing software.

Graphics calculator instructions for Casio fx-9860G Plus, Casio fx-CG20, TI-84 Plus and TI-*n*spire are available from icons located throughout the book.

Fully worked solutions are provided at the back of the text, however students are encouraged to attempt each question before referring to the solution.

It is not our intention to define the course. Teachers are encouraged to use other resources. We have developed this book independently of the International Baccalaureate Organization (IBO) in consultation with experienced teachers of IB Mathematics. The Text is not endorsed by the IBO.

In this changing world of mathematics education, we believe that the contextual approach shown in this book, with associated use of technology, will enhance the students understanding, knowledge and appreciation of mathematics and its universal applications.

We welcome your feedback.

Email: info@haesemathematics.com.au

CTQ PJB

Web: www.haesemathematics.com.au

RCH PMH

ACKNOWLEDGEMENTS

The authors and publishers would like to thank all those teachers who offered advice and encouragement on this book.

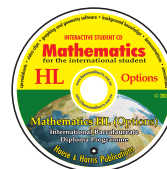
USING THE INTERACTIVE STUDENT CD

The interactive CD is ideal for independent study.

Students can revisit concepts taught in class and undertake their own revision and practice. The CD also has the text of the book, allowing students to leave the textbook at school and keep the CD at home.

By clicking on the relevant icon, a range of interactive features can be accessed:

- ♦ Graphics calculator instructions for the **Casio fx-9860G Plus**, **Casio fx-CG20**, **TI-84 Plus** and the **TI-*n*spire**
- ♦ Interactive links to graphing software



INTERACTIVE
LINK



GRAPHICS
CALCULATOR
INSTRUCTIONS

TABLE OF CONTENTS

SYMBOLS AND NOTATION USED IN THIS BOOK	6
A Expectation algebra	9
B Discrete random variables	26
C Continuous random variables	42
D Probability generating functions	52
E Distributions of the sample mean and the Central Limit Theorem	66
F Point estimation (unbiased estimators and estimates)	82
G Confidence intervals for means	90
H Significance and hypothesis testing	100
I Bivariate Statistics	124
Review set A	146
Review set B	148
Review set C	151
Review set D	153
THEORY OF KNOWLEDGE (The Central Limit Theorem)	157
THEORY OF KNOWLEDGE (Population Parameters)	158
WORKED SOLUTIONS	160
INDEX	207

SYMBOLS AND NOTATION USED IN THIS BOOK

$\approx$	is approximately equal to
$>$	is greater than
$\geq$	is greater than or equal to
$<$	is less than
$\leq$	is less than or equal to
$\{.....\}$	the set of all elements
$\in$	is an element of
$\notin$	is not an element of
$\mathbb{N}$	the set of all natural numbers $\{0, 1, 2, 3,\}$
$\mathbb{Z}$	the set of integers $\{0, \pm 1, \pm 2, \pm 3,\}$
$\mathbb{Q}$	the set of rational numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{Z}^+$	the set of positive integers $\{1, 2, 3,\}$
$\subseteq$	is a subset of
$\subset$	is a proper subset of
$\Rightarrow$	implies that
$\nRightarrow$	does not imply that
$f: A \rightarrow B$	f is a function under which each element of set A has an image in set B
$f: x \mapsto y$	f is a function under which x is mapped to y
$f(x)$	the image of x under the function f
$f \circ g$	or $f(g(x))$ the composite function of f and g
$ x $	the modulus or absolute value of x
$[a, b]$	the closed interval $a \leq x \leq b$
$]a, b[$	the open interval $a < x < b$
u_n	the n th term of a sequence or series with first term u_1
$\{u_n\}$	the sequence with n th term u_n , if first term is u_1
S_n	the sum of the first n terms of a sequence
S_∞	the sum to infinity of a convergent series
$\sum_{i=1}^n u_i$	$u_1 + u_2 + u_3 + + u_n$
$\prod_{i=1}^n u_i$	$u_1 \times u_2 \times u_3 \times \times u_n$
$\lim_{x \rightarrow a} f(x)$	the limit of $f(x)$ as x tends to a
$\lim_{x \rightarrow a+} f(x)$	the limit of $f(x)$ as x tends to a from the positive side of a
$\lim_{x \rightarrow a-} f(x)$	the limit of $f(x)$ as x tends to a from the negative side of a
$\max\{a, b\}$	the maximum value of a or b
$\sum_{n=0}^{\infty} c_n x^n$	the power series whose terms have form $c_n x^n$
$\frac{dy}{dx}$	the derivative of y with respect to x

$f'(x)$	the derivative of $f(x)$ with respect to x
$\frac{d^2y}{dx^2}$	the second derivative of y with respect to x
$f''(x)$	the second derivative of $f(x)$ with respect to x
$\frac{d^ny}{dx^n}$	the n th derivative of y with respect to x
$f^{(n)}(x)$	the n th derivative of $f(x)$ with respect to x
$\int y \, dx$	the indefinite integral of y with respect to x
$\int_a^b y \, dx$	the definite integral of y with respect to x between the limits $x = a$ and $x = b$
e^x	exponential function of x
$\ln x$	the natural logarithm of x
$\sin, \cos, \tan$	the circular functions
$\csc, \sec, \cot$	the reciprocal circular functions
$\arcsin, \arccos, \arctan$	the inverse circular functions
$\binom{n}{r}$	$\frac{n!}{r!(n-r)!}$
$P(A)$	probability of event A
$P(A')$	probability of the event “not A ”
$P(A B)$	probability of the event A given B
$x_1, x_2, \dots$	observations
$P(x)$	probability distribution function $P_x = P(X = x)$ of the discrete random variable X
$f(x)$	probability density function of the continuous random variable X
$F(x)$	cumulative distribution function of the continuous random variable X
$E(X)$	the expected value of the random variable X
$\text{Var}(X)$	the variance of the random variable X
μ	population mean
σ^2	population variance, the value $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$, for a population of size n
σ	population standard deviation
$\bar{x}$	sample mean
s_n^2	sample variance, the value $s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$, from a sample of size n
s_n	standard deviation of the sample of size n
s_{n-1}^2	unbiased estimate of the population variance, the value
	$s_{n-1}^2 = \frac{n}{n-1} s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$, from a sample of size n
$\bar{X}$	the estimator of μ , that is the function $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, where $X_i, i = 1, \dots, n$ are identically distributed independent random variables each with mean μ

S_n^2	the biased estimator of σ^2 , that is the function $S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ where $X_i, i = 1, \dots, n$ are identically distributed independent random variables each with variance σ^2
S_{n-1}^2	the unbiased estimator of σ^2 , that is the function $S_{n-1}^2 = \frac{n}{n-1} S_n^2$
$\text{DU}(n)$	discrete uniform distribution with parameter n
$\text{B}(1, p)$	Bernoulli distribution with parameter p
$\text{B}(n, p)$	binomial distribution with parameters n and p
$\text{Geo}(p)$	geometric distribution with parameter p
$\text{NB}(r, p)$	negative binomial distribution with parameters r and p
$\text{Po}(m)$	Poisson distribution with mean m
$X \sim \text{DU}(n)$	the random variable X has a discrete uniform distribution with parameter n
$X \sim \text{B}(1, p)$	the random variable X has a Bernoulli distribution with parameter p
$X \sim \text{B}(n, p)$	the random variable X has a binomial distribution with parameters n and p
$X \sim \text{Geo}(p)$	the random variable X has a geometric distribution with parameter p
$X \sim \text{NB}(r, p)$	the random variable X has a negative binomial distribution with parameters r and p
$X \sim \text{Po}(m)$	the random variable X has a Poisson distribution with mean m
$\text{U}(a, b)$	continuous uniform distribution with parameters a and b
$\text{Exp}(\lambda)$	exponential distribution with mean $\frac{1}{\lambda}$
$\text{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
ν	number of degrees of freedom
$t(\nu)$	Student's t -distribution with ν degrees of freedom
$X \sim \text{U}(a, b)$	the random variable X has a continuous uniform distribution with parameters a and b
$X \sim \text{Exp}(\lambda)$	the random variable X has an exponential distribution with mean $\frac{1}{\lambda}$
$X \sim \text{N}(\mu, \sigma^2)$	the random variable X has a normal distribution with mean μ and variance σ^2
$T \sim t(\nu)$	the random variable T has the Student's t -distribution with ν degrees of freedom
$G(t)$	the probability generating function $E(t^X)$ for a discrete random variable X which takes values in $\mathbb{N}$
p	depending on the context, a parameter of a distribution, a population proportion, or a p -value in a hypothesis test
$\hat{p}$	a sample proportion
H_0	null hypothesis
H_1	alternative hypothesis
α	significance level or probability of a Type I error
β	probability of a Type II error
$1 - \beta$	power of a hypothesis test
$\text{Cov}(X, Y)$	covariance of random variables X and Y
ρ	product moment correlation coefficient between two random variables
R	the sample product moment correlation coefficient; an estimator of ρ
r	the observed value of R for a given sample of bivariate data; an estimate of ρ

A

EXPECTATION ALGEBRA

A **random variable** can take any one of a set of values from a given domain, according to given probabilities. The domain may be **discrete** or **continuous**.

DISCRETE RANDOM VARIABLES

If X is a **discrete random variable**, then:

- 1 X has possible values $x_1, x_2, x_3, \dots$. To determine the value of X we usually **count**.
- 2 X takes value x_i with probability p_i , where $0 \leq p_i \leq 1$, $i = 1, 2, 3, \dots$, and $\sum p_i = 1$.
- 3 X has a **probability distribution function** (or **probability mass function**) $P(x)$, where $P(x_i) = P(X = x_i) = p_i$, $i = 1, 2, 3, \dots$.
- 4 X has a **cumulative distribution function** (CDF) $F(x)$, where $F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$.

$F(x)$ is the probability that X takes a value less than or equal to x .

Examples of discrete probability distributions and random variables covered in the Core course include Bernoulli, Discrete Uniform, Binomial, and Poisson.

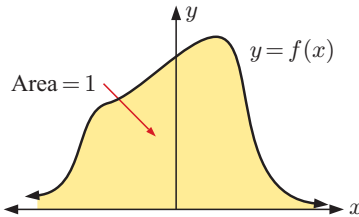
$\sum_{x_i \leq x}$ reads “the sum for all values of x_i less than or equal to x ”.



CONTINUOUS RANDOM VARIABLES

If X is a **continuous random variable**, then:

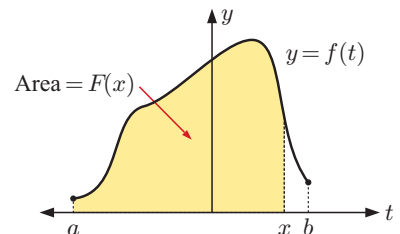
- 1 The possible values of X may be all $x \in \mathbb{R}$, or all real x in some domain $[a, b]$. To determine the value of X we usually **measure**.
- 2 X has a continuous **probability density function** (PDF) $f(x)$, where:



- $f(x) \geq 0$ for all x in the domain of f .
- $\int_{-\infty}^{\infty} f(x) dx = 1$ if the domain of f is $\mathbb{R}$, or $\int_a^b f(x) dx = 1$ if the domain of f is $[a, b]$.

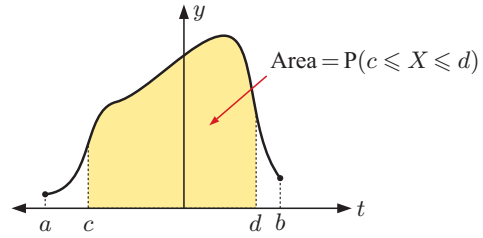
- 3 Suppose f and the PDF for X , have domain $[a, b]$. X has a **cumulative distribution function** (CDF) $F(x)$, where $F(x) = P(X \leq x) = \int_a^x f(t) dt$ for $x \in [a, b]$.

- $F(x)$ is the probability that X takes a value less than or equal to x .
- $F(b) = \int_a^b f(t) dt = 1$



- The probability that X takes a value in the interval $[c, d] \subseteq [a, b]$ is given by

$$\begin{aligned} P(c \leq X \leq d) &= \int_c^d f(t) dt \\ &= F(d) - F(c) \end{aligned}$$



- 4 Since X has infinitely many possible values, the probability that X takes a single value $X = x$ is 0. However, since X is a continuous random variable, for $x \in \mathbb{Z}$ and $x - 0.5 \leq X < x + 0.5$, the value of X will be rounded to the integer x .

Thus, for $x \in \mathbb{Z}$, we define $P(X = x) = P(x - 0.5 \leq X < x + 0.5)$

$$\begin{aligned} &= \int_{x-0.5}^{x+0.5} f(t) dt \\ &= F(x + 0.5) - F(x - 0.5) \end{aligned}$$

Example 1

Given a random variable $X \sim N(7.2, 28)$, find $P(X = 10)$.

$$\begin{aligned} P(X = 10) &= P(9.5 \leq X < 10.5) \\ &\approx 0.0655 \end{aligned}$$

You should recognise the Normal distribution from the Core course.



$$\begin{aligned} \text{For a continuous random variable } X, \quad &P(c \leq X < d) \\ &= P(c \leq X \leq d) \\ &= P(c < X \leq d) \\ &= P(c < X < d) \end{aligned}$$

since the corresponding integrals all define the same area under the curve $y = f(t)$ between $t = c$ and $t = d$.

Examples of continuous probability distributions and random variables covered in the Core course include Continuous Uniform, Exponential, and the Normal distribution.

EXPECTATION

The **mean** or **expected value** or **expectation** $E(X)$ of a random variable X is defined as follows:

- If X is a discrete random variable with set of possible values $x_1, x_2, \dots$ and probability mass function $P(X = x_i) = p_i$, $i = 1, 2, \dots$, $E(X) = \mu = \sum_i x_i P(X = x_i)$

$$= \sum_i x_i p_i, \quad i = 1, 2, \dots$$

- If X is a continuous random variable with probability density function $f(x)$ with domain $[a, b]$,

$$E(X) = \mu = \int_a^b x f(x) dx.$$

Example 2

Find the expectation of the following random variables:

a a discrete random variable X with probability distribution:

x	1	2
$P(X = x)$	$\frac{2}{3}$	$\frac{1}{3}$

b the continuous uniform random variable $X \sim U(1, 4)$.

$$\begin{aligned} \mathbf{a} \quad E(X) &= 1 \times \frac{2}{3} + 2 \times \frac{1}{3} \\ &= \frac{4}{3} \end{aligned}$$

$$\begin{aligned} \mathbf{b} \quad E(X) &= \int_1^4 x \times \frac{1}{3} dx \\ &= \left[\frac{x^2}{6} \right]_1^4 \\ &= \frac{15}{6} \\ &= \frac{5}{2} \end{aligned}$$

Theorem 1

- 1** $E(d) = d$ for any constant d .
- 2** $E(cX + d) = cE(X) + d$ for X a random variable and constants $c, d \in \mathbb{R}$.

Proof:

- 1** A constant d takes value d with probability 1.

$\therefore$ by definition, $E(d) = d \times 1 = d$.

- 2** Discrete Case: X takes value x_i with probability p_i .

Let $U = cX + d$ be a new random variable.

$\therefore U$ takes value $cx_i + d$ with probability p_i , $i = 1, 2, \dots$

$\therefore$ by the definition of $E(U)$,

$$\begin{aligned} E(U) &= E(cX + d) = \sum (cx_i + d) p_i \\ &= \sum (cx_i p_i + d p_i) \\ &= c \sum x_i p_i + d \sum p_i \\ &= cE(X) + d \quad \{\text{since } \sum p_i = 1\} \end{aligned}$$

Continuous Case: If X has PDF $f(x)$ with domain $[a, b]$, then

$$\begin{aligned} E(cX + d) &= \int_a^b (cx + d) f(x) dx \\ &= c \int_a^b x f(x) dx + d \int_a^b f(x) dx \\ &= cE(X) + d \quad \{\text{since } \int_a^b f(x) dx = 1\} \end{aligned}$$

Example 3

Consider a 6-sided unbiased die with face labels: $-2, -2, 0, 1, 1, 2$.

Let X be the discrete random variable with possible values the outcome of a roll of the die.

- a** Find $E(X)$ and $E(X^2)$. **b** Show that $E(X^2) \neq (E(X))^2$.

- a** The discrete random variable X has probability distribution:

x	-2	0	1	2
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$

$$\therefore E(X) = -2 \times \frac{1}{3} + 0 \times \frac{1}{6} + 1 \times \frac{1}{3} + 2 \times \frac{1}{6} = 0$$

Consider the variable X^2 which has possible values 0, 1, and 4.

X^2 has probability distribution:

X^2	0	1	4
Probability	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

$$\begin{aligned} P(X^2 = 4) &= P(X = -2 \text{ or } 2) \\ &= \frac{1}{3} + \frac{1}{6} \\ &= \frac{1}{2} \end{aligned}$$

$$\therefore E(X^2) = 0 \times \frac{1}{6} + 1 \times \frac{1}{3} + 4 \times \frac{1}{2} = \frac{7}{3}$$

- b** $E(X^2) = \frac{7}{3} \neq 0 = (E(X))^2$

In **Example 3** we see that

$$\begin{aligned} E(X^2) &= 0^2 \times \frac{1}{6} + 1^2 \times \frac{1}{3} + 4 \times \left(\frac{1}{3} + \frac{1}{6}\right) \\ &= 0^2 \times \frac{1}{6} + 1^2 \times \frac{1}{3} + (-2)^2 \times \frac{1}{3} + 2^2 \times \frac{1}{6} \\ &= \sum_{i=1}^4 x_i^2 p_i \end{aligned}$$

where X has values $x_1 = -2, x_2 = 0, x_3 = 1, x_4 = 2$ with probabilities $p_1 = \frac{1}{3}, p_2 = \frac{1}{6}, p_3 = \frac{1}{3}, p_4 = \frac{1}{6}$ respectively.

$$\therefore E(g(x)) = \sum g(x_i) p_i, \text{ where } g(X) = X^2.$$

This result generalises for any function g .

Theorem 2

Suppose X is a random variable and g is any function. The random variable $g(X)$ has mean given by:

Discrete case: $E(g(X)) = \sum g(x_i) p_i$

Continuous case: $E(g(X)) = \int g(x) f(x) dx$

Proof for the discrete case:

$g(X)$ has values $g(x_1), g(x_2), \dots$ which are not necessarily distinct.

$g(X)$ takes value $g(x_i)$ with probability $p_i, i = 1, 2, 3, \dots$, or if $g(x) = g(x_1) = \dots = g(x_k)$ then $g(X)$ takes value $g(x)$ with probability $p_1 + p_2 + \dots + p_k$.

Hence, from the definition of expected value, $E(g(x)) = \sum g(x_i) p_i$.

The proof for the continuous case is analogous.



Corollary:

For X a random variable, $E(cg(X) \pm dh(X)) = cE(g(X)) \pm dE(h(X))$
 where c, d are constants and $g(x)$ and $h(x)$ are functions.

Proof for the discrete case:

$$\begin{aligned}
 \text{By Theorem 1 and Theorem 2, } E(cg(X) \pm dh(X)) &= \sum (cg(x_i) \pm dh(x_i)) p_i \\
 &= c \sum g(x_i) p_i \pm d \sum h(x_i) p_i \\
 &= cE(g(X)) \pm dE(h(X))
 \end{aligned}$$

VARIANCE

The **variance** σ^2 , also denoted $\text{Var}(X)$, of a random variable X is
 $\sigma^2 = \text{Var}(X) = E((X - \mu)^2)$, where $\mu = E(X)$.

$\text{Var}(X)$ is the mean of the squared differences of values of X from the mean value of X , $E(X)$.
 $\text{Var}(X)$ is therefore a measure of the spread of the distribution of X .

Theorem 3

For X a random variable, $\text{Var}(X) = E(X^2) - \{E(X)\}^2$.

Proof:

$$\begin{aligned}
 \text{Var}(X) &= E((X - \mu)^2) \\
 &= E(X^2 - 2X\mu + \mu^2) \\
 &= E(X^2) - E(2\mu X) + E(\mu^2) \quad \{\text{by the Corollary to Theorem 2}\} \\
 &= E(X^2) - 2\mu E(X) + \mu^2 \quad \{\text{by Theorem 1}\} \\
 &= E(X^2) - 2\mu^2 + \mu^2 \\
 &= E(X^2) - \mu^2 \\
 &= E(X^2) - \{E(X)\}^2
 \end{aligned}$$

Using **Theorem 3**, we find that:

$$\text{Discrete case: } \text{Var}(X) = \left(\sum x_i^2 p_i \right) - \mu^2$$

$$\text{Continuous case: } \text{Var}(X) = \int x^2 f(x) dx - \mu^2$$

Theorem 4

For X a random variable and constants $c, d \in \mathbb{R}$:

- 1** $\text{Var}(d) = 0$
- 2** $\text{Var}(cX + d) = c^2 \text{Var}(X)$

Proof:

- 1** For any constant d ,

$$\begin{aligned}
 \text{Var}(d) &= E(d^2) - \{E(d)\}^2 \quad \{\text{by Theorem 3}\} \\
 &= d^2 - d^2 \quad \{\text{by Theorem 1 since } d^2, d \text{ are constants}\} \\
 &= 0
 \end{aligned}$$

$$\begin{aligned}
\mathbf{2} \quad \text{Var}(cX + d) &= E((cX + d)^2) - \{E(cX + d)\}^2 \\
&= E(c^2X^2 + 2cdX + d^2) - \{cE(X) + d\}^2 \\
&= c^2E(X^2) + \cancel{2cdE(X)} + E(d^2) - \{c^2(E(X))^2 + \cancel{2cdE(X)} + d^2\} \\
&= c^2E(X^2) + d^2 - c^2(E(X))^2 - d^2 \\
&= c^2\{E(X^2) - \{E(X)\}^2\} \\
&= c^2\text{Var}(X) \quad \{\text{by Theorem 3}\}
\end{aligned}$$

LINEAR TRANSFORMATION OF A SINGLE RANDOM VARIABLE

Let X be a random variable. Let U be a new random variable obtained from X by the linear transformation $U = cX + d$, where c, d are constants.

By **Theorems 1** and **4** we have:

$$E(U) = E(cX + d) = cE(X) + d$$

$$\text{and } \text{Var}(U) = \text{Var}(cX + d) = c^2\text{Var}(X)$$

Theorem 5

Let X be a random variable with mean μ and standard deviation σ .

Then $X^* = \frac{X - \mu}{\sigma}$ is a random variable with $E(X^*) = 0$ and $\text{Var}(X^*) = 1$.

$X^* = \frac{X - \mu}{\sigma}$ is called the **standardised variable** corresponding to X .

Proof:

$$\begin{aligned}
E(X^*) &= E\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) & \text{and} & & \text{Var}(X^*) &= \text{Var}\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) \\
&= \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} & & & &= \frac{1}{\sigma^2}\text{Var}(X) \\
&= \frac{1}{\sigma} \times \mu - \frac{\mu}{\sigma} & & & &= \frac{1}{\sigma^2} \times \sigma^2 \\
&= 0 & & & &= 1
\end{aligned}$$

For example, if $X \sim N(\mu, \sigma^2)$ is a continuous random variable with a normal distribution with mean μ and variance σ^2 , then $Z = \frac{X - \mu}{\sigma}$ has $E(Z) = 0$ and $\text{Var}(Z) = 1$.

$Z \sim N(0, 1)$, where $N(0, 1)$ is the **standard normal distribution**, as studied in the Core course.

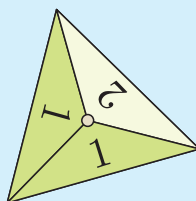
DEFINING NEW RANDOM VARIABLES FROM OLD

In the following example we use two random variables X and Y to define new random variables $3X - 5$, X^2 , $X + Y$, and XY .

We observe how the mean and variance of the new random variables are related to the mean and variance of X and Y , as shown in **Theorems 1 - 4**.

Example 4

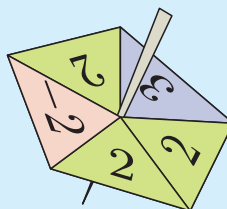
Let X be the random variable with value the outcome of spinning the equilateral triangle spinner:



X and Y are independent random variables.



Let Y be the random variable with value the outcome of spinning the regular pentagon spinner:



- a** Determine the probability distributions for each of:
- i** X and X^2
 - ii** Y and Y^2
 - iii** $3X - 5$ and $(3X - 5)^2$
 - iv** $X + Y$ and $(X + Y)^2$
 - v** XY and $(XY)^2$
- b** Find the mean and variance of:
- i** X
 - ii** Y
 - iii** $3X - 5$
 - iv** $X + Y$
 - v** XY
- c** Verify that:
- i** $\{E(X)\}^2 \neq E(X^2)$
 - ii** $E(3X - 5) = 3E(X) - 5$ and $\text{Var}(3X - 5) = 3^2 \times \text{Var}(X)$
 - iii** $E(X + Y) = E(X) + E(Y)$ and $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
 - iv** $E(XY) = E(X)E(Y)$

- a i** X has probability distribution:

X	1	2
Probability	$\frac{2}{3}$	$\frac{1}{3}$

X has values 1, 2

$\therefore X^2$ has values $1^2 = 1$, $2^2 = 4$

$\therefore X^2$ has probability distribution:

X^2	1	4
Probability	$\frac{2}{3}$	$\frac{1}{3}$

- ii** Y has probability distribution:

Y	-2	2	3
Probability	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$

Y has values -2, 2, 3

$\therefore Y^2$ has values $(-2)^2 = 4$, $2^2 = 4$, $3^2 = 9$

$$P(Y^2 = 9) = P(Y = 3) = \frac{1}{5} \quad \text{and} \quad P(Y^2 = 4) = P(Y = -2 \text{ or } Y = 2)$$

$$= \frac{1}{5} + \frac{3}{5}$$

$$= \frac{4}{5}$$

$\therefore Y^2$ has probability distribution:

Y^2	4	9
Probability	$\frac{4}{5}$	$\frac{1}{5}$

iii

X	1	2
$3X - 5$	-2	1
$(3X - 5)^2$	4	1
Probability	$\frac{2}{3}$	$\frac{1}{3}$

iv The possible outcomes of $X + Y$ are:

		Y			
		X + Y	-2	2	3
X	1	-1	3	4	
	2	0	4	5	

$$\begin{aligned}
 &P((X + Y) = -1) \quad \text{and} \quad P((X + Y) = 4) \\
 &= P(X = 1 \text{ and } Y = -2) \quad = P(X = 2 \text{ and } Y = 2) + P(X = 1 \text{ and } Y = 3) \\
 &= P(X = 1) \times P(Y = -2) \quad = \frac{1}{3} \times \frac{3}{5} + \frac{2}{3} \times \frac{1}{5} \quad \{\text{since } X \text{ and } Y \text{ are independent}\} \\
 &= \frac{2}{3} \times \frac{1}{5} \quad = \frac{5}{15} \\
 &= \frac{2}{15} \quad \text{and similarly for the remaining values.}
 \end{aligned}$$

 $\therefore X + Y$ has probability distribution:

$X + Y$	-1	0	3	4	5
Probability	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{6}{15}$	$\frac{5}{15}$	$\frac{1}{15}$

 $\therefore (X + Y)^2$ has probability distribution:

$(X + Y)^2$	1	0	9	16	25
Probability	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{6}{15}$	$\frac{5}{15}$	$\frac{1}{15}$

v The possible values of XY are:

		Y			
		XY	-2	2	3
X	1	-2	2	3	
	2	-4	4	6	

$$\begin{aligned}
 P(XY = 3) &= P(X = 1 \text{ and } Y = 3) \\
 &= P(X = 1) \times P(Y = 3) \quad \{\text{since } X \text{ and } Y \text{ are independent}\} \\
 &= \frac{2}{3} \times \frac{1}{5} \\
 &= \frac{2}{15} \quad \text{and similarly for the remaining values.}
 \end{aligned}$$

 $\therefore XY$ has probability distribution:

XY	-4	-2	2	3	4	6
Probability	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{6}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{1}{15}$

 $\therefore (XY)^2$ has values $(-4)^2 = 16$, $(-2)^2 = 4$, $2^2 = 4$, $3^2 = 9$, $4^2 = 16$, $6^2 = 36$.

$$\begin{aligned}
 P((XY)^2 = 16) &= P(XY = -4 \text{ or } XY = 4) \\
 &= P(XY = -4) + P(XY = 4) \\
 &= \frac{1}{15} + \frac{3}{15} \\
 &= \frac{4}{15} \quad \text{and similarly for the remaining values.}
 \end{aligned}$$

 $\therefore (XY)^2$ has probability distribution:

$(XY)^2$	4	9	16	36
Probability	$\frac{8}{15}$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{1}{15}$

- b**
- i** $E(X) = 1 \times \frac{2}{3} + 2 \times \frac{1}{3} = \frac{4}{3}$
 $\text{Var}(X) = E(X^2) - \{E(X)\}^2$
 $= \{1 \times \frac{2}{3} + 4 \times \frac{1}{3}\} - \{\frac{4}{3}\}^2$
 $= 2 - \frac{16}{9}$
 $= \frac{2}{9}$
- ii** $E(Y) = -2 \times \frac{1}{5} + 2 \times \frac{3}{5} + 3 \times \frac{1}{5} = \frac{7}{5}$
 $\text{Var}(Y) = E(Y^2) - \{E(Y)\}^2$
 $= \{4 \times \frac{4}{5} + 9 \times \frac{1}{5}\} - \{\frac{7}{5}\}^2$
 $= 5 - \frac{49}{25}$
 $= \frac{76}{25}$
- iii** $E(3X - 5) = -2 \times \frac{2}{3} + 1 \times \frac{1}{3} = -1$
 $\text{Var}(3X - 5) = E((3X - 5)^2) - \{E(3X - 5)\}^2$
 $= \{4 \times \frac{2}{3} + 1 \times \frac{1}{3}\} - \{-1\}^2$
 $= 2$
- iv** $E(X + Y) = -1 \times \frac{2}{15} + 0 \times \frac{1}{15} + 3 \times \frac{6}{15} + 4 \times \frac{5}{15} + 5 \times \frac{1}{15}$
 $= \frac{41}{15}$
 $\text{Var}(X + Y) = E((X + Y)^2) - \{E(X + Y)\}^2$
 $= \{1 \times \frac{2}{15} + 0 \times \frac{1}{15} + 9 \times \frac{6}{15} + 16 \times \frac{5}{15} + 25 \times \frac{1}{15}\} - \{\frac{41}{15}\}^2$
 $= \frac{734}{225}$
- v** $E(XY) = -4 \times \frac{1}{15} - 2 \times \frac{2}{15} + 2 \times \frac{6}{15} + 3 \times \frac{2}{15} + 4 \times \frac{3}{15} + 6 \times \frac{1}{15}$
 $= \frac{28}{15}$
 $\text{Var}(XY) = E((XY)^2) - \{E(XY)\}^2$
 $= \{4 \times \frac{8}{15} + 9 \times \frac{2}{15} + 16 \times \frac{4}{15} + 36 \times \frac{1}{15}\} - \{\frac{28}{15}\}^2$
 $= \frac{1466}{225}$
- c**
- i** $E(X^2) = 1 \times \frac{2}{3} + 4 \times \frac{1}{3} = 2$ and $\{E(X)\}^2 = \{\frac{4}{3}\}^2 = \frac{16}{9} \neq 2$.
Hence $E(X^2) \neq \{E(X)\}^2$
- ii** $E(3X - 5) = -1$ and $3E(X) - 5 = 3 \times \frac{4}{3} - 5 = -1$.
Hence $E(3X - 5) = 3E(X) - 5$.
 $\text{Var}(3X - 5) = 2$ and $3^2 \times \text{Var}(X) = 9 \times \frac{2}{9} = 2$.
Hence $\text{Var}(3X - 5) = 3^2 \text{Var}(X)$.
- iii** $E(X + Y) = \frac{41}{15}$ and $E(X) + E(Y) = \frac{4}{3} + \frac{7}{5} = \frac{41}{15}$.
Hence $E(X + Y) = E(X) + E(Y)$.
 $\text{Var}(X + Y) = \frac{734}{225}$ and $\text{Var}(X) + \text{Var}(Y) = \frac{2}{9} + \frac{76}{25} = \frac{734}{225}$.
Hence $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.
- iv** $E(XY) = \frac{28}{15}$ and $E(X)E(Y) = \frac{4}{3} \times \frac{7}{5} = \frac{28}{15}$.
Hence $E(XY) = E(X)E(Y)$.

We now summarise and prove the results observed in the previous example.

Theorem 6

For X, Y two random variables (either both discrete or both continuous):

- 1** $E(X + Y) = E(X) + E(Y)$
- 2** If X and Y are independent, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
and $E(XY) = E(X)E(Y)$

Proof for the discrete case:

1 $E(X + Y)$

$$= \sum_x \sum_y (x + y) P(X = x \text{ and } Y = y)$$

$$= \sum_x \sum_y x P(X = x \text{ and } Y = y) + \sum_x \sum_y y P(X = x \text{ and } Y = y)$$

$$= \sum_x x \sum_y P(X = x \text{ and } Y = y) + \sum_y y \sum_x P(X = x \text{ and } Y = y)$$

{ x is constant whilst
summing over y .}

{By first changing the order of the
sum, we notice y is constant whilst
summing over x .}

$\sum_x \sum_y$ is used to cover all possible
outcomes for X added to
all possible outcomes for Y .



$$= \sum_x x P(X = x) + \sum_y y P(Y = y)$$

{In a joint probability distribution,

$$P(X = x) = \sum_{\text{all } y} P(X = x \text{ and } Y = y)$$

$$\text{and } P(Y = y) = \sum_{\text{all } x} P(X = x \text{ and } Y = y)$$

This is equivalent to summing a whole row or
whole column in an array of a joint probability
distribution.}

$$= E(X) + E(Y)$$

2 For X and Y independent, $P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$

$$E(XY) = \sum_x \sum_y xy P(X = x \text{ and } Y = y)$$

$$= \sum_x \sum_y xy P(X = x) P(Y = y)$$

$$= \sum_x x P(X = x) \sum_y y P(Y = y)$$

{ x and $P(X = x)$ are constants while summing over y }

$$= \sum_x x P(X = x) E(Y)$$

$$= E(Y) \sum_x x P(X = x) \quad \{E(Y) \text{ is a constant while summing over } x\}$$

$$= E(Y) E(X)$$

$$= E(X) E(Y)$$

$$\text{Var}(X + Y) = E((X + Y)^2) - \{E(X + Y)\}^2$$

$$= E(X^2 + 2XY + Y^2) - \{E(X) + E(Y)\}^2$$

{By part 1 of **Theorem 6**}

$$= E(X^2) + 2E(XY) + E(Y^2)$$

{Corollary to **Theorem 2**}

$$- \{ \{E(X)\}^2 + 2E(X)E(Y) + \{E(Y)\}^2 \}$$

{using $E(XY) = E(X)E(Y)$
from above}

$$= \underbrace{E(X^2) - \{E(X)\}^2}_{\text{Var}(X)} + \underbrace{E(Y^2) - \{E(Y)\}^2}_{\text{Var}(Y)}$$

$$= \text{Var}(X) + \text{Var}(Y)$$

In the following example we show that for X and Y **dependent** random variables either $E(XY) \neq E(X)E(Y)$ or $E(XY) = E(X)E(Y)$, depending on the example.

Example 5

For each of the following distributions for X , let $Y = X^2$ be a random variable dependent on X .

Compare $E(XY)$ with $E(X)E(Y)$.

a

x	-1	0	1
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

b

x	1	2	3
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

a $E(X) = -1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = 0$

Let $Y = X^2$, so Y is dependent on X .

$$\begin{aligned} P(X^2 = 1) &= P(X = 1 \text{ or } X = -1) \\ &= \frac{1}{3} + \frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

$\therefore Y$ has probability distribution:

$Y = X^2$	0	1
Probability	$\frac{1}{3}$	$\frac{2}{3}$

$$\therefore E(Y) = E(X^2) = 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}$$

Consider $XY = X^3$ which has values -1, 0, 1

XY has probability distribution:

$XY = X^3$	-1	0	1
Probability	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$\therefore E(XY) = E(X^3) = -1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = 0$$

$$\text{Hence } E(XY) = 0 \text{ and } E(X)E(Y) = 0 \times \frac{2}{3} = 0$$

Thus $E(XY) = E(X)E(Y)$ even though X and Y are dependent.

b $E(X) = 1 \times \frac{1}{3} + 2 \times \frac{1}{3} + 3 \times \frac{1}{3} = 2$

Let $Y = X^2$ which has values 1, 4, 9

Y	1	4	9
Probability	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$\therefore E(Y) = 1 \times \frac{1}{3} + 4 \times \frac{1}{3} + 9 \times \frac{1}{3} = \frac{14}{3}$$

Consider $XY = X^3$ which has values 1, 8, 27

X^3	1	8	27
Probability	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$\therefore E(X^3) = 1 \times \frac{1}{3} + 8 \times \frac{1}{3} + 27 \times \frac{1}{3} = 12$$

$$\text{Hence } E(XY) = E(X^3) = 12 \text{ but } E(X)E(Y) = 2 \times \frac{14}{3} = \frac{28}{3}$$

$$\therefore E(XY) \neq E(X)E(Y)$$

We now summarise the main results for this section which we obtain from the theorems above and by Mathematical Induction:

Theorem 7

For $X_1, X_2, \dots, X_n$ random variables (either all discrete or all continuous) and $a_1, a_2, \dots, a_n$ constants:

- 1 $E(a_1X_1 \pm a_2X_2 \pm \dots \pm a_nX_n) = a_1E(X_1) \pm a_2E(X_2) \pm \dots \pm a_nE(X_n)$
- 2 If $X_1, X_2, \dots, X_n$ are independent random variables, then
 $E(X_1X_2\dots X_n) = E(X_1)E(X_2)\dots E(X_n)$ and
 $\text{Var}(a_1X_1 \pm a_2X_2 \pm \dots \pm a_nX_n) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + \dots + a_n^2\text{Var}(X_n)$

Proof:

$$\begin{aligned}
 \text{1 By Theorem 1, } E(\pm a_i X_i) &= \pm a_i E(X_i) \\
 \therefore E(a_1 X_1 \pm a_2 X_2 \pm \dots \pm a_{n-1} X_{n-1} \pm a_n X_n) \\
 &= E(a_1 X_1 \pm a_2 X_2 \pm \dots \pm a_{n-1} X_{n-1}) \pm a_n E(X_n) \\
 &= E(a_1 X_1 \pm a_2 X_2 \pm \dots \pm a_{n-2} X_{n-2}) \pm a_{n-1} E(X_{n-1}) \pm a_n E(X_n) \quad \{\text{repeated use of Theorem 6}\} \\
 &\vdots \\
 &= a_1 E(X_1) \pm a_2 E(X_2) \pm \dots \pm a_n E(X_n)
 \end{aligned}$$

$$\begin{aligned}
 \text{2 } E(X_1 X_2 \dots X_n) &= E(X_1 X_2 \dots X_{n-1}) E(X_n) \\
 &= E(X_1 X_2 \dots X_{n-2}) E(X_{n-1}) E(X_n) \quad \{\text{letting } Y = X_1 X_2 \dots X_{n-1}, \\
 &\quad X = X_n\} \\
 &\vdots \\
 &= E(X_1) E(X_2) \dots E(X_n) \quad \{\text{repeated use of Theorem 6}\}
 \end{aligned}$$

By Theorem 4, $\text{Var}(\pm a_i X_i) = a_i^2 \text{Var}(X_i)$

By Theorem 6, letting $X = a_1 X_1 \pm \dots \pm a_{n-1} X_{n-1}$ and $Y = \pm a_n X_n$,

$$\begin{aligned}
 &\text{Var}(a_1 X_1 \pm a_2 X_2 \pm \dots \pm a_{n-1} X_{n-1} \pm a_n X_n) \\
 &= \text{Var}(a_1 X_1 \pm a_2 X_2 \pm \dots \pm a_{n-1} X_{n-1}) + a_n^2 \text{Var}(X_n) \\
 &= \text{Var}(a_1 X_1 \pm \dots \pm a_{n-2} X_{n-2}) + a_{n-1}^2 \text{Var}(X_{n-1}) + a_n^2 \text{Var}(X_n) \quad \{\text{repeated use of Theorem 6}\} \\
 &\vdots \\
 &= a_1^2 \text{Var}(X_1) + \dots + a_n^2 \text{Var}(X_n)
 \end{aligned}$$

CONTINUOUS NORMALLY DISTRIBUTED RANDOM VARIABLES

Theorem 8

Any linear combination of independent continuous normally distributed random variables is itself a continuous and normally distributed random variable.

A linear combination of X_1, X_2, X_3 has the form
 $a_1 X_1 + a_2 X_2 + a_3 X_3$
 where a_1, a_2, a_3 are constants.



For example, if X_1, X_2, X_3 are independent normal random variables then $Y = 2X_1 + 3X_2 - 4X_3$ is a normal random variable.

By **Theorem 7**, $E(Y) = E(2X_1 + 3X_2 - 4X_3)$

$$= 2E(X_1) + 3E(X_2) - 4E(X_3)$$

and $Var(Y) = Var(2X_1 + 3X_2 - 4X_3)$

$$= 4Var(X_1) + 9Var(X_2) + 16Var(X_3)$$

By **Theorem 8**, $Y \sim N(E(Y), Var(Y))$.

Example 6

The weights of male employees in a bank are normally distributed with mean $\mu = 71.5$ kg and standard deviation $\sigma = 7.3$ kg. The bank has an elevator which will carry a maximum load of 444 kg.

- a** Six male employees enter the elevator. Calculate the probability that their combined weight exceeds the maximum load.
- b** If there is to be at most a 0.1% chance of the total weight exceeding 444 kg, recommend the maximum number of males who should use the lift together.

- a** Let the weights of the employees be the independent random variables $X_1, X_2, \dots, X_6$. We are concerned with the sum of their weights $Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$, where $X_i \sim N(71.5, (7.3)^2)$, $i = 1, 2, \dots, 6$.

Now $E(Y) = E(X_1) + E(X_2) + \dots + E(X_6)$

$$= 6 \times 71.5$$

$$= 429 \text{ kg}$$

and $Var(Y)$

$$= Var(X_1) + Var(X_2) + \dots + Var(X_6)$$

$$= 6 \times 7.3^2$$

$$= 319.74$$

$\therefore Y$ is normally distributed with mean 429 kg and variance 319.75 kg^2

$\therefore Y \sim N(429, 319.74)$

Now $P(Y > 444) \approx 0.201$

So, there is a 20.1% chance that their combined weight will exceed 444 kg.

- b** Six men is too many, as there is a 20.1% chance of overload.

Instead we try 5 male employees with total weight $Y = X_1 + X_2 + X_3 + X_4 + X_5$.

$$E(Y) = 5 \times 71.5 \quad Var(Y) = 5 \times 7.3^2$$

$$= 357.5 \text{ kg} \quad \approx 266.45 \text{ kg}^2$$

Now $Y \sim N(357.5, 266.45)$

$\therefore P(Y > 444) \approx 5.82 \times 10^{-8}$

For $n = 5$, there is much less than a 0.1% chance of the total weight exceeding 444 kg. We recommend that a maximum of 5 men use the elevator together.

Example 7

The random variable X has distribution with mean 11 and standard deviation 2. Define three independent random variables $X_1 = 2X$, $X_2 = 4 - 3X$, and $X_3 = 4X + 1$. Find the mean and standard deviation of the random variable $(X_1 + X_2 + X_3)$.

Mean	Variance
$= E(X_1 + X_2 + X_3)$	$= \text{Var}(X_1 + X_2 + X_3)$
$= E(X_1) + E(X_2) + E(X_3)$	$= \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)$
$= 2E(X) + (4 - 3E(X)) + (4E(X) + 1)$	$= 4\text{Var}(X) + 9\text{Var}(X) + 16\text{Var}(X)$
$= 3E(X) + 5$	$= 29\text{Var}(X)$
$= 3(11) + 5$	$= 29 \times 2^2$
$= 38$	$= 116$

$\therefore$ the mean is 38 and the standard deviation is $\sqrt{116} \approx 10.8$.

Example 8

A cereal manufacturer produces packets of cereal in two sizes, small and economy. The amount in each packet is distributed normally and independently with mean and variance as shown in the table.

	Mean (g)	Variance (g^2)
Small	315	4
Economy	950	25

- A packet of each size is selected at random. Find the probability that the economy packet contains less than three times the amount in the small packet.
- One economy and three small packets are selected at random. Find the probability that the economy packet contains less than the total amount in the three small packets.

Let S be the weight of a small packet and E be the weight of an economy packet.
 $\therefore S \sim N(315, 4)$ and $E \sim N(950, 25)$.

- The probability that the economy packet contains less than three times the amount in a small packet, is $P(E < 3S)$ or $P(E - 3S < 0)$.

$$\begin{aligned}
 \text{Now } E(E - 3S) & \quad \text{and} \quad \text{Var}(E - 3S) \\
 &= E(E) - 3E(S) &= \text{Var}(E) + 9\text{Var}(S) \\
 &= 950 - 3 \times 315 &= 25 + 9 \times 4 \\
 &= 5 &= 61
 \end{aligned}$$

$$\therefore E - 3S \sim N(5, 61)$$

$$\therefore P(E - 3S < 0) \approx 0.261$$

- The probability that the economy packet contains less than the total amount in the three small packets is $P(E < S_1 + S_2 + S_3)$ or $P(E - (S_1 + S_2 + S_3) < 0)$, where S_1, S_2, S_3 each have distribution $N(315, 4)$.

$$\begin{aligned}
 \text{Now } E(E - (S_1 + S_2 + S_3)) & \quad \text{and} \quad \text{Var}(E - (S_1 + S_2 + S_3)) \\
 &= E(E) - 3E(S) &= \text{Var}(E) + \text{Var}(S_1) + \text{Var}(S_2) + \text{Var}(S_3) \\
 &= 950 - 3 \times 315 &= 25 + 12 \\
 &= 5 &= 37
 \end{aligned}$$

$$\therefore E - (S_1 + S_2 + S_3) \sim N(5, 37)$$

$$\therefore P(E - (S_1 + S_2 + S_3) < 0) \approx 0.206$$

EXERCISE A

- 1 Given a random variable $X \sim N(5, 22)$, find:

a $P(X = 24)$ b $P(X = 30)$ c $P(X = 11)$

- 2 Find the expectation of the following random variables:

- a a discrete random variable X with probability distribution:

x	-1	1	3
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

- b the continuous uniform random variable $X \sim U(0, 5)$.

- 3 Consider a 6-sided unbiased die with face labels $-3, -1, -1, 2, 2, 3$. Let X be the discrete random variable with possible values the outcome of a roll of the die.

- a Find $E(X)$ and $E(X^2)$. b Show that $E(X^2) \neq (E(X))^2$.

- 4 Let X be the number of heads obtained when an unbiased coin is tossed twice.
Let Y be the result when an ordinary unbiased 6-sided die is rolled once.

- a Write down the probability distributions for each of:

i X and X^2 ii Y and Y^2 iii $X + Y$ and $(X + Y)^2$

iv $4X - 2Y$ and $(4X - 2Y)^2$ v XY and $(XY)^2$

- b Find the mean and variance of:

i X ii Y iii $X + Y$ iv $4X - 2Y$ v XY

- c Verify that:

i $E(X + Y) = E(X) + E(Y)$ and $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$

ii $E(4X - 2Y) = 4E(X) - 2E(Y)$ and $\text{Var}(4X - 2Y) = 16\text{Var}(X) + 4\text{Var}(Y)$

iii $E(XY) = E(X)E(Y)$

- 5 X and Y are two discrete random variables with $E(X) = 3$, $E(Y) = 2$, $\text{Var}(X) = \frac{3}{2}$, and $\text{Var}(Y) = \frac{5}{4}$. Calculate, where possible:

a $E(X + Y)$ b $E(XY)$ c $\text{Var}(2X - 3Y + 6)$ d $E(5XY)$

Hint: You need to consider the two cases: X, Y are independent and X, Y are dependent.

- 6 For each of the following distributions for X , let $Y = X^2$ be a random variable. Hence compare $E(XY)$ with $E(X)E(Y)$.

a

x	0	1	2	3
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

b

x	-1	0	1	2
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

- 7 Consider two independent normally distributed random variables X and Y with means and standard deviations shown.

- a Find the mean and standard deviation of $3X - 2Y$.

- b Find $P(3X - 2Y > 3)$. State clearly any results you use.

	μ	σ
X	3.8	0.323
Y	5.7	1.02

- 8 The marks in an IB Mathematics HL exam are distributed normally with mean μ and standard deviation σ . The cut off score for a 7 is a mark of 80%, and 10% of students get a 7. The cut off score for a 6 is a mark of 65%, and 20% of students get a 6. Find the mean and standard deviation of the marks in this exam.

- 9** The maximum load of a lift is 440 kg. The weights of adults are normally distributed with mean 81 kg and standard deviation 11 kg. The weights of children are normally distributed with mean 48 kg and standard deviation 4 kg.
Find the probability that if the lift contains 4 adults and 3 children then the maximum load will be exceeded. What assumption have you made in your calculation?
- 10** A coffee machine dispenses a cappuccino made up of black coffee distributed normally with mean 120 mL and standard deviation 7 mL, and froth distributed normally with mean 28 mL and standard deviation 4.5 mL.
Each cup is marked to a level of 135.5 mL, and if this is not attained then the customer will receive their cappuccino free of charge.
Determine whether or not the proprietor needs to adjust the settings on her machine if she wishes to give away no more than 1% of cappuccinos free.
- 11** X and Y are independent normal random variables with $X \sim N(-10, 1)$ and $Y \sim N(25, 25)$.
a Find the mean and standard deviation of the random variable $U = 3X + 2Y$.
b Find $P(U < 0)$.
- 12** A drinks manufacturer produces bottles of drink in two sizes, small (S) and large (L). The distributions for the contents are independent, and normally distributed with $S \sim N(280 \text{ mL}, 4 \text{ mL}^2)$ and $L \sim N(575 \text{ mL}, 16 \text{ mL}^2)$.
a A bottle of each size is selected at random. Find the probability that the large bottle contains less than two times the amount in the small bottle.
b One large and two small bottles are selected at random. Find the probability that the large bottle contains less than the total amount in the two small bottles.
- 13** Chocolate bars are produced independently in two sizes, small and large. The amount in each bar is distributed normally and independently as $S \sim N(21, 5)$ and $L \sim N(90, 15)$.
a One of each type of bar is selected at random. Find the probability that the large bar contains more than five times the amount in the small bar.
b One large and five small bars are selected at random. Find the probability that the large bar contains more than the total amount in the five small bars.
- 14** In a gambling game you bet on the outcomes of two spinners. These outcomes are X and Y with the following probability distributions:

x	-3	-2	3	5
$P(X = x)$	0.25	0.25	0.25	0.25

y	-3	2	5
$P(Y = y)$	0.5	0.3	0.2

- a** Briefly explain why these are *well-defined* probability distributions.
- b** Find the mean and standard deviation of each random variable.
- c** Suppose it costs \$1 to get a spinner spun and you receive the dollar value of the outcome. For example, if the result is 3 you win \$3, but if the result is -3 you need to pay an extra \$3. In which game are you likely to achieve a better result? On average, do you expect to win, lose, or break even? Justify your answer.
- d** Comment on the differences in standard deviation.

- e The players are now invited to play a \$1 game using the *sum* of the scores obtained on each of the spinners. For example, if the sum of the spinners is 10, you receive \$10 after paying out \$1. Effectively you win \$9.
- i Copy and complete the table below to show the probability distribution of $X + Y$. A grid may help you do this.

$x + y$	-6	-5	...	10
$P(x + y)$		0.125		

- ii Calculate the mean and standard deviation of the variable $U = X + Y$.
- iii Are you likely to win, lose, or draw in the new game? Justify your answer.

B**DISCRETE RANDOM VARIABLES**

In this section we present important examples of discrete random variables and examine their **cumulative distribution functions (CDF)**.

DISCRETE UNIFORM

A **discrete uniform random variable** X takes n distinct values $x_1, x_2, \dots, x_n$, and the probability mass function is a constant. $P(X = x_i) = \frac{1}{n}$, $i = 1, 2, \dots, n$.

$$X \text{ has CDF } F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i).$$

We write $X \sim \text{DU}(n)$.

For example, suppose an unbiased 6-sided die is labelled $-2, -1, 0, 1, 2, 3$.

The random variable X for the possible outcomes is $X \sim \text{DU}(6)$ since

$$P(X = -2) = P(X = -1) = P(X = 0) = P(X = 1) = P(X = 2) = P(X = 3) = \frac{1}{6}.$$

The probability that the outcome from the roll of this die is less than or equal to 1 is

$$\begin{aligned} F(1) &= P(X \leq 1) \\ &= P(X = -2) + P(X = -1) + P(X = 0) + P(X = 1) \\ &= \frac{4}{6}. \end{aligned}$$

BERNOULLI

A **Bernoulli random variable** X has two possible values 1 ('success') and 0 ('failure').

$$P(X = 1) = p = \text{the probability of success in one trial}$$

$$P(X = 0) = 1 - p = \text{the probability of failure in one trial}$$

where p is a constant, $0 \leq p \leq 1$.

$$X \text{ has CDF } F(x) = \sum_{k=0}^x p^k (1-p)^{1-k},$$

but note that there are only two values: $F(0) = 1 - p$ and $F(1) = 1$.

We write $X \sim \text{B}(1, p)$.

Example 9

Consider the toss of an unbiased coin. Let $X = 1$ if a head is tossed and let $X = 0$ if a tail is the result.

- a** Show that X is a Bernoulli variable.
- b** Suppose that the CDF of X is $F(x)$. Find $F(0)$ and $F(1)$, and interpret their meaning.

- a** $P(X = 1) = \frac{1}{2} = p$
 $P(X = 0) = \frac{1}{2} = 1 - p$
 $\therefore X \sim \text{B}(1, \frac{1}{2})$.

- b** $F(0)$ is the probability of no heads.

$$\begin{aligned} F(0) &= P(\text{tail}) \\ &= P(X = 0) \\ &= \frac{1}{2} \end{aligned}$$

$F(1)$ is the probability of at most one head.

$$\begin{aligned} F(1) &= P(X \leq 1) & \text{or } F(1) &= \sum_{k=0}^1 p^k (1-p)^{1-k} \\ &= P(X = 0) + P(X = 1) & &= p^0 (1-p)^1 + p^1 (1-p)^0 \\ &= \frac{1}{2} + \frac{1}{2} = 1 & &= (1-p) + p \\ & & &= 1 \end{aligned}$$

The probability of “no heads or one head” covers all possibilities and therefore it is a certain event.

BINOMIAL

A **binomial random variable** X has $n + 1$ distinct possible values, $x = 0, 1, 2, \dots, n$ where x is the number of successes in n independent Bernoulli trials $B(1, p)$.

Thus $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2, \dots, n$, where p is a constant, $0 \leq p \leq 1$.

$$\begin{aligned} X \text{ has CDF } F(x) &= P(X \leq x) \\ &= \sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

We write $X \sim B(n, p)$.

When $n = 1$, X is a Bernoulli random variable.



We note from the HL Core course that $E(X) = np$ and $\text{Var}(X) = np(1-p)$.

Example 10

A bag contains 3 red balls and 4 blue balls, all identical in shape.

Consider $n = 8$ trials of randomly choosing a ball from the bag, noting its colour, and then replacing the ball in the bag.

Find the probability of choosing up to 2 red balls.

Let X be the number of red balls selected in 8 such (independent) trials.

Then $P(X = x) = \binom{8}{x} \left(\frac{3}{7}\right)^x \left(\frac{4}{7}\right)^{8-x}$, $x = 0, 1, 2, \dots, 8$ and $X \sim B(8, \frac{3}{7})$.

$$\begin{aligned} \text{The probability of choosing up to 2 red balls is } F(2) &= \sum_{k=0}^2 \binom{8}{k} \left(\frac{3}{7}\right)^k \left(\frac{4}{7}\right)^{8-k} \\ &= \left(\frac{4}{7}\right)^8 + 8 \left(\frac{3}{7}\right)^1 \left(\frac{4}{7}\right)^7 + \binom{8}{2} \left(\frac{3}{7}\right)^2 \left(\frac{4}{7}\right)^6 \\ &\approx 0.259 \end{aligned}$$

EXERCISE B.1

- 1 The discrete random variable X is such that $P(X = x) = k$, for $x = 5, 10, 15, 20, 25, 30$. Find:

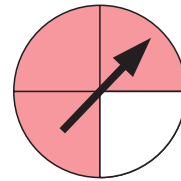
- a the probability distribution of X
- b μ , the expected value of X
- c $P(X < \mu)$
- d σ , the standard deviation of X .

For more questions on the uniform and binomial distributions, see the Core course **Chapter 25**.



- 2 Consider the spinner alongside. Let $X = 1$ if the outcome is red, and $X = 0$ if the outcome is white.

- a Show that X is a Bernoulli variable.
- b Suppose that the CDF of X is $F(x)$. Find $F(0)$ and $F(1)$, and interpret their meaning.



- 3 Consider the random variable X such that $X \sim B(7, p)$, $p < 0.5$, and $P(X = 4) = 0.09724$. Find $P(X = 2)$.
- 4 In Statsville, the probability that it will rain on any given day in August is 0.35. Calculate the probability that in a given week in August in Statsville, it will rain on:
- a exactly 3 days
 - b at least 3 days
 - c at most 3 days
 - d exactly 3 days in succession.

State any assumptions made in your calculations.

- 5 A box contains a very large number of red and blue pens. The probability that a randomly selected pen is blue, is 0.8. How many pens would you need to select to be more than 90% certain of picking at least one red pen? State any assumptions made in your calculations.
- 6 A satellite relies on solar cells for its operation and will be powered provided at least one of its cells is working. Solar cells operate independently of each other, and the probability that an individual cell fails within one year is 0.7.
- a For a satellite with 15 solar cells, find the probability that:
 - i all 15 cells fail within one year
 - ii the satellite is still operating at the end of one year.
 - b For a satellite with n solar cells, find the probability that it is still operating at the end of one year.
 - c Hence, find the smallest number of cells required so that the probability of the satellite still operating at the end of one year is at least 0.98.
- 7 Seventy percent (70%) of the mail to ETECH Couriers is addressed to the Accounts Department.
- a In a batch of 20 letters, what is the probability that there will be at least 11 letters to the Accounts Department?
 - b On average 70 letters arrive each day. Find the mean and standard deviation of the number of letters to the Accounts Department each day.

- 8 The table gives information about the destination and type of parcels handled by ETECH Couriers.

- a Find the probability that a parcel is being sent interstate given that it is priority paid.

Hint: Use **Bayes theorem**, see HL Core text, **Chapter 24**.

Destination		Priority	Standard
Local	40%	70%	30%
Country	20%	45%	55%
Interstate	25%	70%	30%
International	15%	40%	60%

- b If two standard parcels are selected, find the probability that exactly one will be leaving the state, either interstate or international.
- 9 At a school fete fundraiser, an unbiased spinning wheel has numbers 1 to 50 inclusive.
- a Find the probability of getting a multiple of 7 in one spin of the wheel.
- b If the wheel is spun 500 times during the day, what is the likelihood of getting a multiple of 7 more than 15% of the time?
- c Suppose 20 people play each time the wheel is spun. When a multiple of 7 comes up, \$5 is paid to players, but when it does not the players must pay \$1.



- i How much would the wheel be expected to make or lose for the school if it was spun 500 times?
- ii Find the probability that the school will lose money if the wheel is spun 500 times during the day.

GEOMETRIC

Suppose a sports magazine gives away photographs of famous football players. 15 photographs are randomly placed in every 100 magazines.

Let X be the number of magazines you purchase before you get a photograph.

$$P(X = 1) = P(\text{the first magazine contains a photo}) = 0.15$$

$$P(X = 2) = P(\text{the second magazine contains a photo}) = 0.15 \times 0.85$$

$$P(X = 3) = P(\text{the third magazine contains a photo}) = 0.15 \times (0.85)^2$$

$$\therefore P(X = x) = 0.15 \times (0.85)^{x-1} \text{ for } x = 1, 2, 3, 4, \dots$$

This is an example of a *geometric* distribution.

If X is the number of independent Bernoulli trials $B(1, p)$, $0 < p \leq 1$, needed to obtain a successful outcome, then X is a **geometric discrete random variable** and has probability mass function

$$P(X = x) = p(1 - p)^{x-1} \text{ where } x = 1, 2, 3, 4, \dots$$

$$\text{The CDF is } F(x) = P(X \leq x) = \sum_{k=1}^x p(1 - p)^{k-1} \text{ for } x = 1, 2, 3, 4, \dots$$

We write $X \sim \text{Geo}(p)$.

Example 11

In a spinning wheel game with numbers 1 to 50 on the wheel, you win if you get a multiple of 7. Assuming the game is fair, find the probability that you will win:

- a after exactly four games
- b after no more than three games
- c after more than three games.

Let X be the number of games played until you win.

$$X \sim \text{Geo}(p) \text{ where } p = \frac{7}{50} = 0.14 \text{ and } 1 - p = 0.86$$

$$\begin{aligned} \mathbf{a} \quad & P(X = 4) \\ &= p(1 - p)^3 \\ &= 0.14 \times (0.86)^3 \\ &\approx 0.0890 \end{aligned}$$

$$\begin{aligned} &= P(\text{wins after no more than three games}) \\ &= P(X \leq 3) \\ &= p + p(1-p) + p(1-p)^2 \\ &= p(1 + (1-p) + (1-p)^2) \\ &= 0.14(1 + 0.86 + 0.86^2) \\ &\approx 0.364 \end{aligned}$$

$$\begin{aligned} \text{or } & \text{P(wins after no more than three games)} \\ &= \text{P}(X \leq 3) \\ &= 1 - \text{P}(\text{does not win in any of the first three games}) \\ &= 1 - (1 - p)^3 \\ &= 1 - 0.86^3 \\ &\approx 0.364 \end{aligned}$$

- $$\begin{aligned} \text{P(wins after more than three games)} &= \text{P}(X > 3) \\ &= 1 - \text{P}(X \leq 3) \\ &\approx 1 - 0.364 \quad \{\text{from } \mathbf{b}\} \\ &\approx 0.636 \end{aligned}$$

Theorem 9

Suppose $X \sim \text{Geo}(p)$, $0 < p < 1$, is a geometric discrete random variable.

$$\mathbf{1} \quad F(x) = P(X \leq x) = \sum_{k=1}^x p(1-p)^{k-1} = 1 - (1-p)^x$$

2 The modal value of X is $x = 1$.

Proof:

$$\begin{aligned} \mathbf{1} \quad \sum_{k=1}^x p(1-p)^{k-1} &= p \sum_{k=1}^x (1-p)^{k-1} \\ &= p\{1 + (1-p) + (1-p)^2 + \dots + (1-p)^{x-1}\} \\ &= p \left\{ \frac{1 - (1-p)^x}{p} \right\} \quad \{\text{since we have a finite GS}\} \\ &= 1 - (1-p)^x \end{aligned}$$

2 $P(X = 1) = p$
 $P(X = x) = p(1 - p)^{x-1} \leq p$ for $x \geq 1$, since $0 < p < 1$.
 $\therefore P(X = 1) = p$ is the highest probability in the distribution.
 $\therefore x = 1$ is the mode of the distribution of $X \sim \text{Geo}(p)$.

NEGATIVE BINOMIAL (PASCAL'S DISTRIBUTION)

If X is the number of independent Bernoulli trials $B(1, p)$, $0 < p < 1$, required for r successes then X has a **negative binomial** distribution.

We observe that if $r = 1$ then the negative binomial distribution is a geometric distribution.

Example 12

In grand slam mens tennis, the player who wins a match is the first player to win 3 sets. Suppose that $P(\text{Novak beats Rafael in one set}) = 0.55$. Find the probability that when Novak plays Rafael in the grand slam event:

- a Novak wins the match in three sets
- b Novak wins the match in four sets
- c Novak wins the match in five sets
- d Rafael wins the match.

Let X be the number of sets played until Novak wins.

- a $P(X = 3)$
 $= (0.55)^3$
 ≈ 0.166
- b $P(X = 4)$
 $= P(\text{RNNN or NRNN or NNRN})$
 $= 3 \times 0.55^3 \times 0.45^1 \approx 0.225$
- c $P(X = 5)$
 $= P(\text{RRNNN or RNRNN or RNNRN or NRRNN or NRNRN or NNRRN})$
 $= 6 \times 0.55^3 \times 0.45^2$
 ≈ 0.202
- d $P(\text{Rafael wins the match})$
 $= 1 - P(\text{Novak wins the match})$
 $= 1 - (0.55^3 + 3 \times 0.55^3 \times 0.45 + 6 \times 0.55^3 \times 0.45^2)$
 ≈ 0.407

Examining **b** from the above **Example 12**, we notice that

$$P(X = 4) = P(\text{Novak wins 2 of the first 3 and wins the 4th}) = \underbrace{\binom{3}{2} (0.55)^2 (0.45)^1}_{\text{binomial}} \times 0.55$$

Generalising,

$$\begin{aligned} P(X = x) &= P(r - 1 \text{ successes in } x - 1 \text{ independent trials and success in the last trial}) \\ &= \binom{x-1}{r-1} p^{r-1} (1-p)^{x-r} \times p \\ &= \binom{x-1}{r-1} p^r (1-p)^{x-r} \end{aligned}$$

In repeated independent Bernoulli trials, $B(1, p)$, $0 < p < 1$, where p is the probability of success in each trial, let X denote the number of trials needed to gain r successes.

X has a **negative binomial distribution** with probability mass function

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad r \geq 1, \text{ where } x = r, r+1, r+2, \dots$$

The CDF is $F(x) = P(X \leq x) = \sum_{k=r}^x \binom{k-1}{r-1} p^r (1-p)^{k-r}$ where $1 \leq r \leq k \leq x$.

We write $X \sim \text{NB}(r, p)$ and say X is a **negative binomial random variable**.

Example 13

On each point in a badminton set, Dan has probability 0.52 of beating Chong. Suppose they play one set to 21 points. Find the probability that Dan wins the set 21-19.

Let X be the number of points played until Dan wins.

$$\therefore X \sim \text{NB}(21, 0.52)$$

$$\begin{aligned}\therefore P(\text{Dan wins 21-19}) &= P(X = 40) \\ &= \binom{39}{20} 0.52^{21} 0.48^{19} \\ &\approx 0.0658\end{aligned}$$

EXERCISE B.2

- 1 X is a discrete random variable where $X \sim \text{Geo}(0.25)$. Without using technology, calculate:
 - a $P(X = 4)$
 - b $P(X \leq 2)$
 - c $P(X > 3)$
- 2 Suppose $X \sim \text{Geo}(p)$, $0 < p < 1$. Show that the probability distribution is well defined, so $\sum_{i=1}^{\infty} P(X = i) = 1$.
- 3 In a game of ten-pin bowling, Xu has a 29% chance of getting a 'strike' with every bowl he attempts at all ten pins.
 - a Find the probability of Xu getting a 'strike' after exactly 4 bowls.
 - b Find the probability that Xu will take 7 bowls to score 3 'strikes'.
- 4 Suppose $X \sim \text{Geo}(p)$ and that the probability that the first success is obtained on the 3rd attempt is 0.023 987. If $p > 0.5$, find $P(X \geq 3)$.
- 5 In any game of squash, Paul has a 65% chance of beating Eva. To win a match in squash, a player must win three games.
 - a State the nature of the distribution in a squash match.
 - b Find the probability that Eva beats Paul by 3 games to 1.
 - c Find the probability that Eva beats Paul in a match of squash.
- 6 From past experience, Fred has a 72% chance of driving a golf ball in excess of 230 m. Find the probability that:
 - a Fred needs 5 drives in order to hit one in excess of 230 m.
 - b Fred needs 12 drives in order to exceed 230 m 4 times.
- 7 At a luxury ski resort in Switzerland, the probability that snow will fall on any given day in the snow season is 0.15.
 - a If the snow season begins on November 1st, find the probability that the first snow will fall on November 15.
 - b Given that no snow fell during November, a tourist decides to wait no longer to book a holiday. On December 1st, the tourist decides to book for the earliest date for which the probability that the first snow will have fallen is greater than 0.85. Find the exact date of the booking.



- 8 In a board game for four players, the players must each roll two fair dice in turn to get a difference of “no more than 3”, before they can move their marker on the board.
- Find the probability of getting a difference of “no more than 3” when rolling two unbiased dice.
 - Player 1 rolls the dice first. Find the probability that player 1 is the first to move his counter, and this happens on his second roll.

POISSON

The **Poisson distribution** was studied in **Chapter 25** of the Core text.

X is a **Poisson random variable** if $E(X) = \text{Var}(X) = m > 0$ and X has probability mass function $P(X = x) = \frac{m^x e^{-m}}{x!}$, where $x = 0, 1, 2, 3, \dots$.

The CDF of X is $F(x) = P(X \leq x) = \sum_{k=0}^x \frac{m^k e^{-m}}{k!}$ for $x = 0, 1, 2, 3, \dots$.

We write $X \sim \text{Po}(m)$.

0! = 1



We have seen that a binomial random variable is used to describe the number of successes in a certain number of independent Bernoulli trials.

A Poisson random variable X can be interpreted as the number of successes (or occurrences) in an interval of given, specific length, when the following conditions hold:

- The average number $E(X) = m$, of occurrences is known and is constant for all intervals of the given, specific length.
- The number of occurrences in such intervals are independent when the intervals are disjoint.
- The probability of success in any given trial is small.

Examples of random variables which can be modelled with a Poisson distribution include:

- the number of typing errors per page in a book.
- the number of telephone calls per hour received by an office.

Example 14

Let X be the number of patients that arrive at a hospital emergency room per hour. Patients arrive at random and the average number of patients per hour is constant.

- Explain why X is a Poisson random variable.
- Suppose we know that $3\text{Var}(X) = [E(X)]^2 - 4$.
 - Find the mean of X .
 - Find $P(X \leq 4)$.
- Let Y be the number of patients admitted to the hospital Intensive Care Unit each hour. Suppose it also has a Poisson distribution with $\text{Var}(Y) = 3$, and that Y is independent of X .
 - Find $E(X + Y)$ and $\text{Var}(X + Y)$.
 - What do you suspect about the distribution of $X + Y$?
- Let U be the random variable defined by $U = X - Y$.
 - Find the mean and variance of U .
 - Comment on the distribution of U .

- a** X is a Poisson random variable as the average number of patients arriving at random per hour is constant. We assume that the numbers of patients arriving each hour are independent.
- b** **i** Since $E(X) = \text{Var}(X) = m$, we find $3m = m^2 - 4$
 $\therefore m^2 - 3m - 4 = 0$
 $\therefore (m - 4)(m + 1) = 0$
 $\therefore m = 4 \text{ or } -1$
 But $m > 0$, so $m = 4$
- ii** $X \sim \text{Po}(4)$. $P(X \leq 4) \approx 0.629$
- c** **i** $E(X + Y)$ $\text{Var}(X + Y)$
 $= E(X) + E(Y)$ $= \text{Var}(X) + \text{Var}(Y)$
 $= 4 + 3$ $\{E(Y) = \text{Var}(Y) = 3\}$ $= 4 + 3$
 $= 7$ $= 7$
- ii** Since $E(X + Y) = \text{Var}(X + Y)$, we suspect $X + Y$ is a Poisson random variable, and that $X + Y \sim \text{Po}(7)$.
- d** **i** $E(U)$ $\text{Var}(U)$
 $= E(X - Y)$ $= \text{Var}(X - Y)$
 $= E(X) - E(Y)$ $= \text{Var}(X) + \text{Var}(Y)$
 $= 4 - 3$ $= 4 + 3$
 $= 1$ $= 7$
- ii** Since $E(U) \neq \text{Var}(U)$, the variable $U = X - Y$ cannot be Poisson.

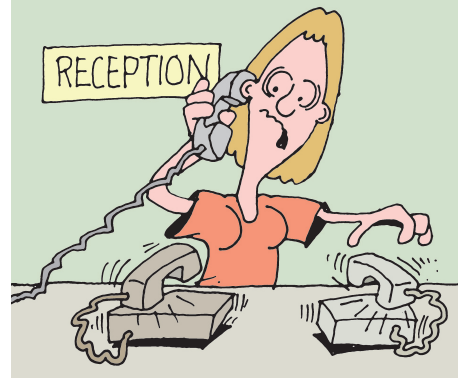
We will prove in the next exercise that the sum of two independent Poisson random variables is itself a Poisson random variable.



EXERCISE B.3

- 1** X is a discrete random variable such that $X \sim \text{Po}(\mu)$ and $P(X = 2) = P(X = 0) + 2P(X = 1)$.
- a** Find the value of μ . **b** Hence, evaluate $P(1 \leq X \leq 5)$.
- 2** Let X be the number of emergency calls made to the police per hour. The calls arrive at random, and the average number of calls per hour is constant.
- a** Explain why X is a Poisson random variable.
b Suppose we know that $2\text{Var}(X) = [E(X)]^2 - 15$.
- i** Find the mean of X . **ii** Find $P(X \leq 3)$.
- 3** In a mining process, the workers regularly use chains of length 50 metres. It is known that chains from a particular manufacturer have faults at the average rate of 1 per every kilometre of chain.
- a** Find the probability that there will be:
- i** no faults in a 50 metre length of chain
ii at most two faults in the 50 metre length of chain.
- b** A chain is considered 'safe' if there is at least a 99.5% chance there will be no more than 1 fault in 50 m of chain. Should the chains from this manufacturer be considered 'safe'?

- 4 A receptionist in an international school receives on average five internal calls per 20 minutes and ten external calls per half hour.
- Calculate the probability that the receptionist will receive exactly three calls in five minutes.
 - On average, how many calls will the receptionist receive every five minutes? Give your answer to the nearest integer.
 - Find the probability that the receptionist receives more than five calls in:
 - 5 minutes
 - 7 minutes.



Example 15

Consider two independent Poisson random variables X and Y , both with mean m .
Prove that $X + Y$ is a Poisson random variable with mean $2m$.

Since X and Y are two Poisson random variables with mean m , $P(X = x) = \frac{m^x e^{-m}}{x!}$ and $P(Y = y) = \frac{m^y e^{-m}}{y!}$.

Since X and Y are independent,

$$P(X = x \text{ and } Y = y) = P(X = x) \times P(Y = y)$$

$$\begin{aligned} \therefore P(X + Y = k) &= \sum_{i=0}^k (P(X = i \text{ and } Y = k - i)) \\ &= \sum_{i=0}^k P(X = i) \times P(Y = k - i) \\ &= \sum_{i=0}^k \frac{m^i e^{-m}}{i!} \times \frac{m^{k-i} e^{-m}}{(k-i)!} \\ &= \frac{e^{-2m}}{k!} \sum_{i=0}^k \frac{k!}{i!(k-i)!} m^i m^{k-i} \\ &= \frac{e^{-2m}}{k!} (2m)^k \quad \{\text{Binomial theorem}\} \end{aligned}$$

$\therefore X + Y$ is a Poisson random variable with mean $2m$.

- 5 a Let $X \sim \text{Po}(m_1)$ and $Y \sim \text{Po}(m_2)$ be two independent Poisson random variables.
Prove that $X + Y$ has probability mass function

$$P(X + Y = k) = \frac{(m_1 + m_2)^k e^{-(m_1 + m_2)}}{k!}, \quad k = 0, 1, 2, \dots$$

and thereby prove that $X + Y$ is a Poisson random variable with distribution $\text{Po}(m_1 + m_2)$.

- b Prove by induction that the sum of n independent Poisson random variables $X_1, X_2, \dots, X_n$, with corresponding means $m_1, m_2, \dots, m_n$ respectively, is a Poisson random variable with distribution $\text{Po}(m_1 + m_2 + \dots + m_n)$.

THE MEAN AND VARIANCE OF DISCRETE RANDOM VARIABLES

We have seen that to calculate the mean and variance of a discrete random variable we use:

- the **mean** $E(X) = \mu = \sum x_i p_i$
- the **variance** $\text{Var}(X) = \sigma^2 = \sum (x_i - \mu)^2 p_i$
 $= \sum x_i^2 p_i - \mu^2$
 $= E(X^2) - \{E(X)\}^2$

Example 16

Consider $X \sim \text{DU}(n)$ where $X = 1, 2, \dots, n$, which is a special case of a discrete uniform random variable.

Given that $1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$ for all n in $\mathbb{Z}^+$, show that $E(X) = \frac{n+1}{2}$ and $\text{Var}(X) = \frac{n^2 - 1}{12}$.

$$\begin{aligned}
 E(X) &= \sum x_i p_i \\
 &= 1 \left(\frac{1}{n}\right) + 2 \left(\frac{1}{n}\right) + 3 \left(\frac{1}{n}\right) + \dots + n \left(\frac{1}{n}\right) \\
 &= \frac{1}{n} (1 + 2 + 3 + 4 + \dots + n) \quad \text{where } 1 + 2 + 3 + \dots + n \text{ is an} \\
 &\quad \text{arithmetic series with } u_1 = 1 \text{ and } d = 1 \\
 &= \frac{1}{n} \left[\frac{n}{2} (2u_1 + (n-1)d) \right] \\
 &= \frac{1}{2} [2 + (n-1)] \\
 &= \frac{n+1}{2}
 \end{aligned}$$

$$\begin{aligned}
 \text{Var}(X) &= \sum x_i^2 p_i - \mu^2 \\
 &= 1^2 \left(\frac{1}{n}\right) + 2^2 \left(\frac{1}{n}\right) + 3^2 \left(\frac{1}{n}\right) + \dots + n^2 \left(\frac{1}{n}\right) - \left(\frac{n+1}{2}\right)^2 \\
 &= \frac{1}{n} (1^2 + 2^2 + 3^2 + \dots + n^2) - \frac{(n+1)^2}{4} \\
 &= \frac{1}{n} \left[\frac{n(n+1)(2n+1)}{6} \right] - \frac{(n+1)^2}{4} \\
 &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\
 &= (n+1) \left[\frac{2n+1}{6} - \frac{n+1}{4} \right] \\
 &= (n+1) \left[\frac{4n+2}{12} - \frac{3n+3}{12} \right] \\
 &= (n+1) \left[\frac{n-1}{12} \right] \\
 &= \frac{n^2 - 1}{12}
 \end{aligned}$$

For $X \sim \text{DU}(n)$ in general, X takes values $x_1, x_2, \dots, x_n$ where $P(X = x_i) = p_i = \frac{1}{n}$ for all $i = 1, \dots, n$. The mean and variance are calculated as:

$$\begin{aligned} E(X) &= \sum_{i=1}^n x_i p_i = \frac{1}{n} \sum_{i=1}^n x_i \\ \text{Var}(X) &= \sum_{i=1}^n x_i^2 p_i - \{E(X)\}^2 \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{1}{n^2} \left\{ \sum_{i=1}^n x_i \right\}^2. \end{aligned}$$

By similar direct calculation of the mean and variance for discrete distributions, the results in the following table can be obtained. We will return to them later in the course when we study **probability generating functions**.

Distribution	Notation	Probability mass function	Mean	Variance
Bernoulli	$X \sim B(1, p)$	$p^x(1-p)^{1-x}$ for $x = 0, 1$	p	$p(1-p)$
Binomial	$X \sim B(n, p)$	$\binom{n}{x} p^x(1-p)^{n-x}$ for $x = 0, 1, \dots, n$	np	$np(1-p)$
Poisson	$X \sim \text{Po}(m)$	$\frac{m^x e^{-m}}{x!}$ for $x = 0, 1, \dots$	m	m
Geometric	$X \sim \text{Geo}(p)$	pq^{x-1} where $q = 1-p$, for $x = 1, 2, \dots$	$\frac{1}{p}$	$\frac{q}{p^2}$
Negative binomial (Pascal's)	$X \sim \text{NB}(r, p)$	$\binom{x-1}{r-1} p^r q^{x-r}$ where $q = 1-p$, for $x = r, r+1, \dots$	$\frac{r}{p}$	$\frac{rq}{p^2}$
Discrete uniform (special case)	$X \sim \text{DU}(n)$	$\frac{1}{n}$ for $x = 1, \dots, n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$

Example 17

- Prove that $x \binom{n}{x} = n \binom{n-1}{x-1}$.
- Hence prove that for a Binomial random variable, the mean is equal to np .

a Proof:

$$\begin{aligned} \text{LHS} &= x \binom{n}{x} & \text{RHS} &= n \binom{n-1}{x-1} \\ &= x \times \frac{n!}{(n-x)!x!} & &= n \times \frac{(n-1)!}{(n-x)!(x-1)!} \\ &= \frac{n!}{(n-x)!(x-1)!} & &= \frac{n!}{(n-x)!(x-1)!} \\ \therefore \text{LHS} &= \text{RHS as required} \end{aligned}$$

b If $X \sim B(n, p)$ then $P(x) = \binom{n}{x} p^x q^{n-x}$ where $q = 1 - p$.

$$\begin{aligned}
 \therefore \mu = E(X) &= \sum_{x=0}^n xP(x) \\
 &= \sum_{x=0}^n x \binom{n}{x} p^x q^{n-x} \\
 &= \sum_{x=1}^n x \binom{n}{x} p^x q^{n-x} && \{\text{since when } x = 0, \text{ the term is } 0\} \\
 &= \sum_{x=1}^n n \binom{n-1}{x-1} p^x q^{n-x} && \{\text{using the above result}\} \\
 &= np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} q^{n-x} \\
 &= np \sum_{r=0}^{n-1} \binom{n-1}{r} p^r q^{n-(r+1)} && \{\text{replacing } x-1 \text{ by } r\} \\
 &= np \sum_{r=0}^{n-1} \binom{n-1}{r} p^r q^{(n-1)-r} \\
 &= np(p+q)^{n-1} && \{\text{Binomial theorem}\} \\
 &= np \times 1 \\
 &= np
 \end{aligned}$$

THE POISSON APPROXIMATION TO THE BINOMIAL DISTRIBUTION

In the following example we observe how, in certain cases, a binomial random variable can be approximated by a Poisson random variable.

Example 18

Sheep are transported to the city using big trucks which carry 500 sheep at a time. On average, 0.8% of the sheep have to be removed on arrival because of illness.

- a** Describe the nature of the random variable X , which indicates the number of ill sheep on arrival.
- b** State the mean and variance of X .
- c** Find the probability that on a truck carrying 500 sheep:
 - i** exactly three are ill on arrival
 - ii** at least four are ill on arrival.
- d** By inspection of your answer to **b**, comment as to what other type of random variable X may approximate.
- e** Use the approximation from **d** to repeat the calculations in **c**. Hence verify the validity of the approximation.

- a** X is a binomial random variable and $X \sim B(500, 0.008)$
- b** $\mu = np = 500 \times 0.008 = 4$ $\sigma^2 = np(1-p) = 4 \times 0.992 \approx 3.97$
- c**
 - i** $P(X = 3) \approx 0.196$
 - ii** $P(\text{at least 4 are ill})$
 $= P(X \geq 4)$
 $= 1 - P(X \leq 3)$
 ≈ 0.567

- d** Using **b**, $\mu \approx \sigma^2$ which suggests we may approximate X using a Poisson distribution. In particular, X is approximately distributed as $\text{Po}(4)$.
- | | | | |
|----------|-------------------|---------------------|---|
| e | $P(X = 3)$ | $P(X \geq 4)$ | These results are excellent approximations to those in c . |
| | ≈ 0.195 ✓ | $= 1 - P(X \leq 3)$ | |
| | | ≈ 0.567 ✓ | |

The previous example prompts the question: Under what conditions can a binomial random variable be approximated by a Poisson random variable?

Let $X \sim B(n, p)$ be a binomial random variable.

X has probability mass function $P(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ where $x = 0, 1, 2, \dots, n$ and $E(X) = np$.

Suppose $np = m$, a constant.

$$\therefore p = \frac{m}{n}$$

For m, x fixed constants:

$$\begin{aligned} P(x) &= \binom{n}{x} \left(\frac{m}{n}\right)^x \left(1 - \frac{m}{n}\right)^{n-x} \\ &= \frac{n(n-1)(n-2) \dots (n-x+1)}{x!} \frac{m^x}{n^x} \left(1 - \frac{m}{n}\right)^{n-x} \\ &= \frac{m^x}{x!} \left\{ \left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \left(\frac{n-2}{n}\right) \dots \left(\frac{n-x+1}{n}\right) \right\} \left(1 - \frac{m}{n}\right)^{n-x} \\ &= \frac{m^x}{x!} \left\{ 1 \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{(x-1)}{n}\right) \right\} \left(1 - \frac{m}{n}\right)^n \left(1 - \frac{m}{n}\right)^{-x} \end{aligned}$$

For a Poisson distribution, the probability p of success in a given trial is small. If we consider $p \rightarrow 0$, we require $n \rightarrow \infty$ in order to keep $np = m$ constant.

$$\therefore \left\{ 1 \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{(x-1)}{n}\right) \right\} \rightarrow 1,$$

$$\left(1 - \frac{m}{n}\right)^{-x} \rightarrow 1,$$

$$\text{and } \left(1 + \frac{-m}{n}\right)^n \rightarrow e^{-m} \quad \{\text{refer to Calculus Option}\}$$

$$\therefore P(x) \rightarrow \frac{m^x e^{-m}}{x!}$$

Thus the binomial distribution approaches a Poisson distribution when $p \rightarrow 0$ and $n \rightarrow \infty$.

For $X \sim B(n, p)$ a binomial random variable, X can be approximated by the Poisson random variable $Y \sim \text{Po}(np)$ provided n is large enough and p is small enough. In general, Y is a reasonable approximation provided $n \geq 50$ and $p \leq 0.1$.

Where appropriate in the following exercise, clearly state the type of discrete distribution used as well as answering the question.

EXERCISE B.4

- 1 Suppose $X \sim \text{Geo}(0.333)$. Find:
 - a the mean of X
 - b the standard deviation of X .
- 2 In a game of ten-pin bowling, Win has a 25% chance of a strike in each frame.
 - a Find the average number of bowls Win requires to get a strike.
 - b What is the average number of bowls Win will need to get two strikes?
- 3 A dart player has a 5% chance of hitting the bullseye with every attempt. Find:
 - a the expected number of throws for this player to get a bullseye
 - b the standard deviation in throws for this player to get a bullseye.
- 4 A spinning wheel has the numbers 1 to 40 inclusive on it. Assuming that the wheel is unbiased, find the mean and standard deviation of all the possible scores when the wheel is spun.
- 5 In an average working week, an office confectionary dispenser breaks down six times. Assume the working week is Monday to Saturday, with each day including the same number of working hours. Which of the following is most likely to occur?
 - A The machine breaks down three times during the week.
 - B The machine breaks down once on Saturday.
 - C The machine breaks down less than seventeen times in a 4 week period.
- 6 

In a World Series contest between the Redsox and the Yankees, the first team to win four games is declared world champion. Recent evidence suggests that the Redsox have a 53% chance of beating the Yankees in any game. Find the probability that:

 - a the Yankees will beat the Redsox in exactly five games
 - b the Yankees will beat the Redsox in exactly seven games
 - c the Redsox will be declared world champions.
 - d On average, how many games would it take for the Redsox to win four games against the Yankees? Comment on your result.
- 7 During the busiest period on the internet, you have a 62% chance of connecting to a government website. If you do not get through, you keep trying until you do connect. Let X be the number of times you have to try in order to get through.
 - a Stating any necessary assumptions, identify the nature of the random variable X .
 - b Find $P(X \geq 3)$.
 - c Find the mean and standard deviation of the random variable X .
- 8 A large aeroplane has 250 passenger seats. From years of business experience, the airline has found that on average 3.75% of travellers who have bought tickets do not arrive for any given flight. The airline sells 255 tickets for this large aeroplane on a particular flight. Let X be the number of ticket holders who do not arrive for the flight.
 - a State the distribution of X .
 - b Calculate the probability that more than 250 ticket holders will arrive for the flight.
 - c Calculate the probability that there will be empty seats on this flight.
 - d For the variable X , calculate the:
 - i mean
 - ii variance.

- e** Hence use a suitable approximation for X to calculate the probability that:
- more than 250 ticket holders will arrive for the flight
 - there will be empty seats on this flight.
- f** Use your answers to determine whether the approximation was a good one.
- 9** It costs €15 to play a game where you have to randomly select a marble from ten differently marked marbles in a barrel. The marbles are marked 10 cents, 20 cents, 30 cents, 40 cents, 50 cents, 60 cents, 70 cents, €15, €30, and €100, and you receive the marked amount in return for playing the game.
- Define a random variable X which is the outcome of selecting a marble from the barrel.
 - Find $E(X)$ and $\text{Var}(X)$.
 - Briefly explain why you cannot use the rules given for $\text{DU}(n)$ to find the answers to **b** above.
 - The people who run the game expect to make a profit but want to encourage people to play by not charging too much.
 - Find, to the nearest 10 cents, the smallest amount they need to charge to still expect to make a profit.
 - Find the expected return to the organisers if they charge €16 per game, and a total of 1000 games are played in one day.
- 10** A person raising funds for cancer research telephones people at random asking for a donation. From past experience, he has a 1 in 8 chance of being successful.
- Describe the random variable X that indicates the number of calls made before he is successful and someone makes a donation. State one assumption made in your answer.
 - Find the average number of calls required for success, and the standard deviation of the number of calls for success.
 - Find the probability that it takes less than five calls to be successful.
- 11** The probability that I dial a wrong number is 0.005 when I make a telephone call. In a typical week I will make 75 telephone calls.
- Describe the distribution of the random variable T that indicates the number of times I dial a wrong number in a week.
 - In a given week, find the probability that:
 - I dial no wrong numbers, $P(T = 0)$
 - I dial more than two wrong numbers, $P(T > 2)$.
 - Find $E(T)$ and $\text{Var}(T)$. Comment on your results.
 - By approximating T with a Poisson distribution, again find the probability that in a given week:
 - I dial no wrong numbers
 - I dial more than two wrong numbers.
 - Discuss your results in **b** and **d**.
- 12**
- Consider $1 + q + q^2 + q^3 + \dots$ where $0 < q < 1$.
 - Find the sum to infinity of this series.
 - Hence, show that $1 + 2q + 3q^2 + 4q^3 + \dots = \sum_{x=1}^{\infty} xq^{x-1} = \frac{1}{(1-q)^2}$ for $0 < q < 1$.
 - If $X \sim \text{Geo}(p)$, prove that $E(X) = \frac{1}{p}$.

C

CONTINUOUS RANDOM VARIABLES

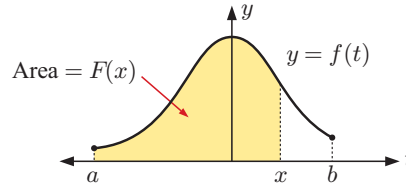
In this section we present important examples of continuous random variables and examine their **cumulative distribution functions (CDF)**. Refer to **Chapter 26** of the Core text to revise the properties of continuous random variables.

In **Section A** we saw that a continuous random variable X with **probability distribution function** $f(x)$ with domain $[a, b]$, has **cumulative distribution function** $F(x)$, where

$$F(x) = P(X \leq x)$$

$$= \int_a^x f(t) dt$$

= area under the curve $y = f(t)$
between $t = a$ and $t = x$.

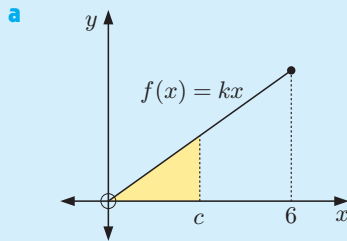


Depending on the form of the function $f(t)$, this area can sometimes be found using simple methods, for example, by finding the area of a triangle or rectangle.

Example 19

The continuous random variable X has PDF $f(x) = kx$, $0 \leq x \leq 6$, where k is a constant.

Find: **a** k **b** the tenth percentile of the random variable X .



Since $\int_0^6 f(x) dx = 1$,

$$\int_0^6 kx dx = 1$$

$$\therefore k \left[\frac{x^2}{2} \right]_0^6 = 1$$

$$\therefore k(18 - 0) = 1$$

$$\therefore k = \frac{1}{18}$$

We could use the area of a triangle formula instead of integrating.



b We need to find c such that $P(X < c) = 0.10$

$$\therefore \frac{1}{2} \times c \times \frac{c}{18} = 0.1 \quad \{\text{The area of the shaded triangle}\}$$

$$\therefore c^2 = 3.6$$

$$\therefore c \approx 1.90 \quad \{\text{as } c > 0\}$$

The 10th percentile ≈ 1.90

THE MEAN AND VARIANCE OF A CONTINUOUS RANDOM VARIABLE

From **Section A** we have the following formulae for calculating the mean and variance of a continuous random variable X :

- $E(X) = \mu = \int x f(x) dx$
- $\text{Var}(X) = \sigma^2 = \int (x - \mu)^2 f(x) dx = E((X - \mu)^2)$

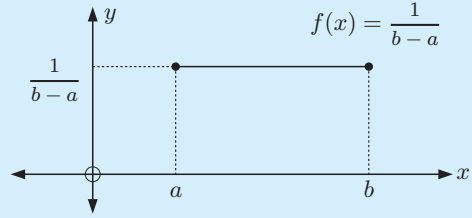
$$= \int x^2 f(x) dx - \mu^2 = E(X^2) - \mu^2$$

CONTINUOUS UNIFORM

A continuous uniform random variable X has:

- domain $[a, b]$, where a, b are constants. The possible values for X are all x such that $a \leq x \leq b$.
- PDF $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$.

Its graph a horizontal line segment as shown.

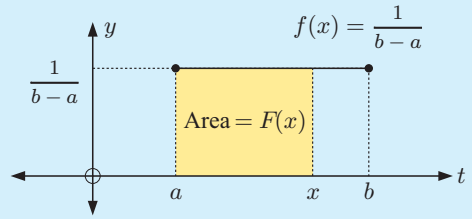


- CDF $F(x) = P(X \leq x) = \int_a^x f(t) dt$

$$= \int_a^x \frac{1}{b-a} dt$$

$$= \frac{x-a}{b-a}$$

= the area of the rectangle shown.



We write $X \sim U(a, b)$.

Example 20

If $X \sim U(a, b)$ is a continuous uniform random variable, show that:

a $\mu = \frac{a+b}{2}$

b $\text{Var}(X) = \frac{(b-a)^2}{12}$

Since $X \sim U(a, b)$, its PDF is $f(x) = \frac{1}{b-a}$, $a \leq x \leq b$.

a $\mu = E(X)$

$$\begin{aligned} &= \int_a^b \frac{x}{b-a} dx \\ &= \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b \\ &= \frac{\frac{b^2}{2} - \frac{a^2}{2}}{b-a} \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{(b+a)(b-a)}{2(b-a)} \\ &= \frac{a+b}{2} \end{aligned}$$

b $\sigma^2 = \text{Var}(X) = E(X^2) - \mu^2$

$$\begin{aligned} &= \int_a^b \frac{x^2}{b-a} dx - \left(\frac{a+b}{2} \right)^2 \\ &= \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b - \left(\frac{a+b}{2} \right)^2 \\ &= \frac{\frac{b^3}{3} - \frac{a^3}{3}}{b-a} - \left(\frac{a+b}{2} \right)^2 \\ &= \frac{b^3 - a^3}{3(b-a)} - \left(\frac{a+b}{2} \right)^2 \\ &= \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)} - \frac{a^2 + 2ab + b^2}{4} \\ &= \frac{4b^2 + 4ab + 4a^2}{12} - \frac{3a^2 + 6ab + 3b^2}{12} \\ &= \frac{b^2 - 2ab + a^2}{12} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

Example 21

The error, in seconds, made by an amateur timekeeper at an athletics meeting may be modelled by the random variable X with probability density function

$$f(x) = \begin{cases} 2.5 & -0.1 \leq x \leq 0.3 \\ 0 & \text{otherwise.} \end{cases}$$

Find the probability that:

- a an error is positive b the magnitude of an error exceeds 0.1 seconds
- c the magnitude of an error is less than 0.2 seconds.

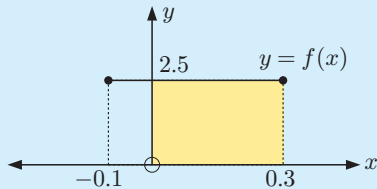
We have $X \sim U(-0.1, 0.3)$ with PDF $f(x) = 2.5$ on $-0.1 \leq x \leq 0.3$.

a $P(X > 0)$

$$= P(0 < X < 0.3)$$

$$= 0.3 \times 2.5 \quad \{\text{Area of the given rectangle}\}$$

$$= 0.75$$



b $P(\text{magnitude} > 0.1)$

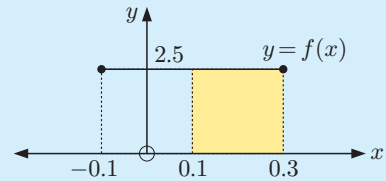
$$= P(|X| > 0.1)$$

$$= P(X > 0.1 \text{ or } X < -0.1)$$

$$= P(X > 0.1)$$

$$= 0.2 \times 2.5 \quad \{\text{Area of rectangle}\}$$

$$= 0.5$$



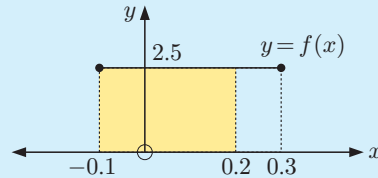
c $P(\text{magnitude} < 0.2) = P(|X| < 0.2)$

$$= P(-0.2 < X < 0.2)$$

$$= P(-0.1 < X < 0.2)$$

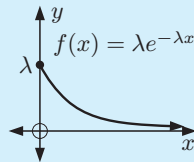
$$= (0.2 - (-0.1)) \times 2.5$$

$$= 0.75$$

**EXPONENTIAL**

A continuous exponential random variable X has:

- domain $[0, \infty[$. The possible values for X are all $x \geq 0$.
- PDF $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, where $\lambda > 0$ is a constant.



- CDF $F(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda t} dt$

$$= [-e^{-\lambda t}]_0^x$$

$$= 1 - e^{-\lambda x}.$$

We write $X \sim \text{Exp}(\lambda)$.

- $\lambda > 0$ is required since $f(x) \geq 0$ is necessary and $e^{-\lambda x} > 0$ for all x .
- $f(x)$ is decreasing for all $x > 0$. This is observed from the graph $y = f(x)$ or algebraically since $f'(x) = -\lambda^2 e^{-\lambda x} < 0$ for all $x > 0$.
- Since $f(x)$ is a PDF, $\int_0^\infty f(x) dx = 1$. Hence $\int_0^\infty \lambda e^{-\lambda t} dt = 1$.
- From the graph of $y = f(x)$ we observe that as $x \rightarrow \infty$, $f(x) \rightarrow 0$ from above.
- Most of the area under the graph of $y = f(x)$ corresponds to low values of x .

For a continuous exponential random variable, $X \sim \text{Exp}(\lambda)$, $E(X) = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.

Example 22

a Show that $f(x)$ is a well-defined PDF. **b** Find $E(X)$ and $\text{Var}(X)$.
c Find the median and modal values of X .

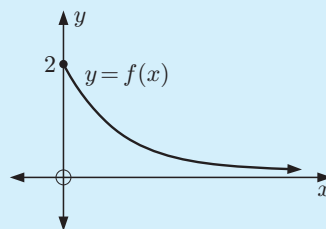
a Now $\int_0^{\infty} f(x) dx = \int_0^{\infty} 2e^{-2x} dx$ and $f(x) = 2e^{-2x} \geq 0$
 since $e^{-2x} \geq 0$ for all $x \in [0, \infty[$.

$$\begin{aligned} &= \lim_{t \rightarrow \infty} \left[\frac{2e^{-2x}}{-2} \right]_0^t \\ &= \lim_{t \rightarrow \infty} [-e^{-2x}]_0^t \\ &= \lim_{t \rightarrow \infty} (1 - e^{-2t}) \\ &= 1 - 0 \\ &= 1 \end{aligned}$$

b Since $X \sim \text{Exp}(2)$, $E(X) = \frac{1}{\lambda} = \frac{1}{2}$ and $\text{Var}(X) = \frac{1}{\lambda^2} = \frac{1}{4}$.

• If the median is m , we need to find m such that $\int_0^m 2e^{-2x} dx = 0.5$

$$\begin{aligned}\therefore \left[\left(\frac{1}{-2} \right) 2e^{-2x} \right]_0^m &= 0.5 \\ \therefore [-e^{-2x}]_0^m &= 0.5 \\ \therefore -e^{-2m} - (-1) &= 0.5 \\ \therefore e^{-2m} &= 0.5 \\ \therefore e^{2m} &= 2 \quad \{\text{reciprocals}\} \\ \therefore 2m &= \ln 2 \\ \therefore m &= \frac{1}{2} \ln 2 \approx 0.347 \\ \therefore \text{the median} &\approx 0.347\end{aligned}$$



$\therefore$ the mode = 0.

Example 23

- a** Find the 80th percentile of the continuous exponential random variable X with PDF $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$, giving your answer in terms of λ .
- b** If $\lambda > 4$, find possible values for the 80th percentile. Comment on your answer.

- a** We need to find c such that

$$\int_0^c \lambda e^{-\lambda t} dt = 0.80$$

$$\therefore \lambda \int_0^c e^{-\lambda t} dt = 0.8$$

$$\therefore \lambda \left[\frac{e^{-\lambda t}}{-\lambda} \right]_0^c = 0.8$$

$$\therefore -[e^{-\lambda c} - e^0] = 0.8$$

$$\therefore e^{-\lambda c} - 1 = -0.8$$

$$\therefore e^{-\lambda c} = 0.2$$

$$\therefore e^{\lambda c} = 5$$

$$\therefore \lambda c = \ln 5$$

$$\therefore c = \frac{\ln 5}{\lambda}$$

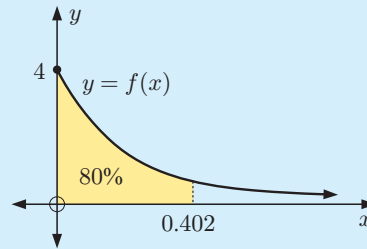
$$\therefore \text{the 80th percentile is } \frac{\ln 5}{\lambda}.$$

- b** If $\lambda > 4$, $\frac{1}{\lambda} < \frac{1}{4}$

$$\therefore \text{the 80th percentile} < \frac{\ln 5}{4} \approx 0.402$$

So, for $\lambda > 4$, 80% of the scores are less than 0.402.

80% of the area under the curve lies in $[0, 0.402]$, which is a very small interval compared with $[0, \infty[$.



Suppose we are given the CDF of a continuous random variable. We can find its PDF using the Fundamental Theorem of Calculus, since the PDF $f(x)$ will also be a continuous function. In particular:

If the CDF is $F(x) = \int_a^x f(t) dt$ then its PDF is given by $f(x) = F'(x)$.

Example 24

Find the PDF of the random variable with CDF $F(x) = \int_0^x \lambda e^{-\lambda t} dt$.

$$\begin{aligned} f(x) &= F'(x) = \frac{d}{dx} \int_0^x \lambda e^{-\lambda t} dt, \quad x \geq 0 \\ &= \frac{d}{dx} \left[\frac{\lambda e^{-\lambda t}}{-\lambda} \right]_0^x \\ &= \frac{d}{dx} [-e^{-\lambda t}]_0^x \\ &= \frac{d}{dx} (-e^{-\lambda x} - (-1)) \\ &= -e^{-\lambda x}(-\lambda) + 0 \end{aligned}$$

$$\therefore \text{the PDF is } f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

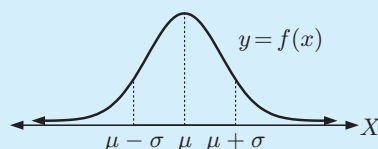
As expected from the Fundamental Theorem of Calculus, $f(x)$ is simply the integrand function.



NORMAL

A continuous normal random variable X has:

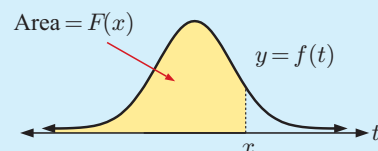
- domain $\mathbb{R}$, so X may take any real value $x \in \mathbb{R}$.
- PDF $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $x \in \mathbb{R}$,
where μ, σ are constants and $\sigma > 0$.



- CDF $F(x) = P(X \leq x)$

$$= \int_{-\infty}^x f(t) dt$$

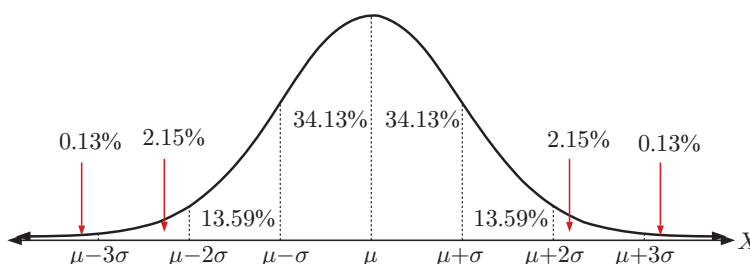
$$= \text{area under } y = f(t) \text{ on }]-\infty, x].$$



We write $X \sim N(\mu, \sigma^2)$.

We note that:

- X has mean $E(X) = \mu$ and variance $\text{Var}(X) = \sigma^2$.
- The normal curve is bell-shaped with the percentages within its portions as shown. Refer to the Core text **Chapter 26** for more information.



- $Z = \frac{X - \mu}{\sigma}$ is the **standard normal random variable**, and $Z \sim N(0, 1)$

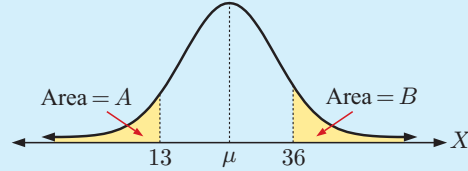
This transformation is useful when determining an unknown mean or standard deviation. Conversion to Z -scores is also very important for understanding the theory behind **confidence intervals** and **hypothesis testing** which are dealt with later in this topic.

SUMMARY OF CONTINUOUS DISTRIBUTIONS

Distribution	Notation	Probability density function	Mean	Variance
Uniform	$X \sim U(a, b)$	$\frac{1}{b-a}, a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$X \sim \text{Exp}(\lambda)$	$\lambda e^{-\lambda x}, x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	$X \sim N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2

Example 25

Consider the random variable $X \sim N(\mu, \sigma^2)$.
Find the mean and standard deviation given that
area $A = 0.115\,06$ and area $B = 0.135\,66$.



$$P(X < 13) = 0.115\,06$$

$$\therefore P\left(\frac{X - \mu}{\sigma} < \frac{13 - \mu}{\sigma}\right) = 0.115\,06$$

$$\therefore P\left(Z < \frac{13 - \mu}{\sigma}\right) = 0.115\,06$$

$$\therefore \frac{13 - \mu}{\sigma} \approx -1.2$$

$$\therefore \mu - 1.2\sigma = 13 \dots (1)$$

$$(2) - (1) \text{ gives } 2.3\sigma = 23$$

$$\therefore \sigma = 10$$

$$\text{Substituting in (1) gives } \mu - 12 = 13$$

$$\therefore \mu = 25$$

$\therefore$ the mean is 25 and the standard deviation is 10.

$$\text{Also, } P(X > 36) = 0.135\,66$$

$$\therefore P\left(Z > \frac{36 - \mu}{\sigma}\right) = 0.135\,66$$

$$\therefore \frac{36 - \mu}{\sigma} \approx 1.1$$

$$\therefore \mu + 1.1\sigma = 36 \dots (2)$$

THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

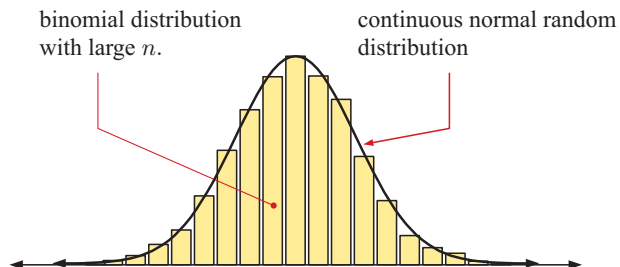
Suppose $X \sim B(n, p)$ is a binomial random variable. For sufficiently large n , we can approximate X (discrete) by X_c (continuous) where $X_c \sim N(np, npq)$ and $q = 1 - p$.

We will prove this result later in the topic.

A useful rule to follow is: If $np > 5$ and $nq > 5$, then any value of X_c is a reasonable approximation for a value of X , provided we allow for a **correction of continuity** (see below).



This can be observed by drawing column graphs for binomial distributions for different values of n and p . When n , p , and q satisfy the above, the column graph has an approximate bell-shape like the PDF of a normal distribution. The greater the values of np and nq , the closer this approximates the graph of the normal distribution.



CORRECTION FOR CONTINUITY

For X_c a continuous random variable, $P(X_c = a) = 0$ for a any constant, but for a an integer we define $P(X_c = a) = P(a - 0.5 \leq X_c < a + 0.5)$, due to rounding on the real number line.

Thus when approximating the discrete binomial random variable $X \sim B(n, p)$ with the continuous normal random variable $X_c \sim N(np, np(1-p))$ we approximate $P(X = 7)$ with $P(6.5 \leq X_c < 7.5)$.

↑
discrete

↑
continuous

Also, for $X \leq 7$ we use $X_c < 7.5$, and for $X \geq 7$ we use $X_c \geq 6.5$.

This is called a **correction for continuity**.

Example 26

Consider the binomial random variable $X \sim B(15, 0.4)$.

- a** Find: **i** $E(X)$ **ii** $\text{Var}(X)$.
- b** Find: **i** $P(X \leq 7)$ **ii** $P(3 \leq X \leq 12)$.
- c** Approximate X with an appropriate continuous normal random variable X_c .
- d** Find: **i** $P(X_c \leq 7)$ **ii** $P(3 \leq X_c \leq 12)$.
Compare your answers with **b**.
- e** Find: **i** $P(X_c < 7.5)$ **ii** $P(2.5 \leq X_c < 12.5)$.
Again, compare your answers with **b**.
- f** Which of the approximations **c** or **d** are better? Explain your answer.

a **i** $E(X) = \mu = np$
 $\therefore E(X) = 15 \times 0.4$
 $= 6$

ii $\text{Var}(X) = \sigma^2 = npq$
 $\therefore \text{Var}(X) = 6 \times 0.6$
 $= 3.6$

b **i** $P(X \leq 7) \approx 0.787$

ii $P(3 \leq X \leq 12)$
 $= P(X \leq 12) - P(X \leq 2)$
 ≈ 0.973

c Using a normal approximation, X is approximately distributed by $X_c \sim N(6, 3.6)$.

d **i** $P(X_c \leq 7) \approx 0.701$

ii $P(3 \leq X_c \leq 12) \approx 0.942$

These answers are not really close to those in **b**.

e Using the normal approximation $X_c \sim N(6, 3.6)$

i $P(X_c < 7.5) \approx 0.785$

ii $P(2.5 \leq X_c < 12.5) \approx 0.967$

These results are very close to the actual values.

f The approximations in **e** are much closer. There has been a correction for continuity because the binomial distribution is *discrete* and the normal distribution is *continuous*.

EXERCISE C

Where appropriate in the following exercise, clearly state the type of discrete or continuous distribution used as well as answering the question.

- 1 The continuous random variable T has the probability density function

$$f(t) = \begin{cases} \frac{1}{2\pi} & -\pi \leq t \leq \pi \\ 0 & \text{otherwise.} \end{cases} \quad \text{Find the mean and standard deviation of } T.$$

- 2 The Australian football Grand Final is held annually on the last Saturday in September. With approximately 100 000 in attendance each year, ticket sales are heavily in demand upon release. Let X be the random variable which gives the time (in hours) required for a successful purchase of a Grand Final ticket after their release. The median value of X is 10 hours.

- a Give reasons why X could best be modelled by a continuous exponential random variable.
- b Find the value of λ in the PDF for the exponential random variable X .
- c Hence, find the probability of a Grand Final ticket being purchased after 3 or more days.
- d Find the average time before a Grand Final ticket is purchased.

- 3 Find the mean and standard deviation of a normal random variable X , given that $P(X > 13) = 0.4529$ and $P(X > 28) = 0.1573$.

- 4 Consider the continuous probability density function $f(x) = \begin{cases} 0, & x < 0 \\ 6 - 18x, & 0 \leq x \leq k \\ 0, & x > k. \end{cases}$

Find:

- a the value of k
 - b the mean and standard deviation of the distribution.
- 5 It is known that 41% of a population support the Environment Party. A random sample of 180 people is selected from the population. Suppose X is the random variable giving the number in the sample who support the Environment Party.
- a State the distribution of X .
 - b Find $E(X)$ and $\text{Var}(X)$.
 - c Find $P(X \geq 58)$.
 - d State a suitable approximation for the random variable X and use it to recalculate c. Comment on your answer.
- 6 When typing a document, trainee typists make on average 2.5 mistakes per page. The mistakes on any one page are made independently of any other page. Suppose X represents the number of mistakes made on one page, and Y represents the number of mistakes made in a 52-page document.
- a State the distributions of X and Y . You may assume that the sum of n independent Poisson random variables, each with mean m , is itself a Poisson random variable with mean mn .
 - b Rana is a trainee typist. Find the probability that Rana will make:
 - i more than 2 mistakes on a randomly chosen page
 - ii more than 104 mistakes in a 52-page document.
 - c Now assume that X and Y can be approximated by normal random variables with the same means and variances as found above. Use the normal approximations to estimate the probabilities in b. Comment on your answers.
- 7 The continuous random variable X has PDF $f(x) = \frac{2}{5}$ where $1 \leq x \leq k$.
- a Find the value of k , and state the distribution of X .
 - b Find $P(1.7 \leq X \leq 3.2)$.
 - c Find $E(X)$ and $\text{Var}(X)$.

- 8 The continuous random variable X is uniformly distributed over the interval $a < x < b$. The 30th percentile is 3 and the 90th percentile is 12. Find:
- a the values of a and b
 - b the PDF of X
 - c $P(5 < X < 9)$
 - d the CDF of X .
- 9
- a If the random variable $T \sim N(7, 36)$, find $P(|T - 6| < 2.3)$.
 - b Four random observations of T are made. Find the probability that exactly two of the observations will lie in the interval $|T - 6| < 2.3$.
- 10 A continuous random variable X has PDF $f(x) = \frac{1}{2}e^{-\frac{1}{2}x}$ for $x \geq 0$.
- a Show that $X \sim \text{Exp}(0.5)$.
 - b Hence find:
 - i μ_X
 - ii σ_X
 - iii the median of X
 - iv the 90th percentile of X .
 - c Use the CDF for the exponential variable to find, correct to 4 decimal places:
 - i $P(X \leq 1)$
 - ii $P(0.4 \leq X \leq 2)$
- 11 The **exponential** probability density function of random variable X is defined as $f(x) = ae^{-ax}$ for $a > 0$ and $x \in [0, \infty[$.
- a On the same set of axes, graph $y = f(x)$ for $a = 1, 2$, and 3 .
 - b Prove that $f(x)$ is a well defined PDF.
 - c Use integration by parts to show that:
 - i $\int axe^{-ax} dx = -e^{-ax} \left(x + \frac{1}{a} \right) + \text{constant}$
 - ii $\int ax^2e^{-ax} dx = -e^{-ax} \left(x^2 + \frac{2x}{a} + \frac{2}{a^2} \right) + \text{constant}$
 - d Show that the mean and variance of the negative exponential variable X are $\frac{1}{a}$ and $\frac{1}{a^2}$ respectively.

Here we prove
Theorem 10.



D

PROBABILITY GENERATING FUNCTIONS

In this section we consider discrete random variables which take values in $\mathbb{N} = \{0, 1, 2, \dots\}$.

First we state some key results required for this section:

1 Finite geometric series (GS)

For $x \in \mathbb{R}$, $x \neq 1$, and $n \in \mathbb{N}$, $\sum_{i=0}^n x^i = 1 + x + x^2 + \dots + x^n = \frac{1 - x^{n+1}}{1 - x}$.

2 Sum of an infinite geometric series (GS)

The infinite sum $\sum_{i=0}^{\infty} x^i = 1 + x + x^2 + x^3 + \dots$ is finite (or convergent) if and only if $|x| < 1$.

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1 - x} \quad \text{if and only if } |x| < 1.$$

3 Binomial formula

For real constants x, y and for $n \in \mathbb{N}$:

$$(x + y)^n = \sum_{i=0}^n \binom{n}{i} x^i y^{n-i} = y^n + \binom{n}{1} xy^{n-1} + \binom{n}{2} x^2 y^{n-2} + \dots + \binom{n}{n-1} x^{n-1} y + x^n$$

4 Exponential series

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \quad \text{for all } x \in \mathbb{R} \quad \{\text{Result from the Calculus Option}\}$$

$$5 \quad \lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = e^a \quad \text{for all } a \in \mathbb{R} \quad \{\text{Result from HL Core}\}$$

6 Binomial series for $r \in \mathbb{Z}^+$ and $|x| < 1$

$$\frac{1}{(1 - x)^r} = \sum_{i=0}^{\infty} x^i (-1)^i \frac{(-r)(-r-1)\dots(-r-(i-1))}{i!} \quad \{\text{Result from the Calculus Option}\}$$

7 Summation identities

$$\sum_{i=1}^n i = 1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^n i^2 = 1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

PROBABILITY GENERATING FUNCTIONS

Let X be a discrete random variable which takes values in $\mathbb{N} = \{0, 1, 2, 3, \dots\}$, and such that $P(X = k) = p_k$, for $k \in \mathbb{N}$.

The **probability generating function** (PGF), $G(t)$, for X is

$$G(t) = E(t^X) = \sum_{k=0}^{\infty} p_k t^k$$

$$= p_0 + p_1 t + p_2 t^2 + \dots \quad \text{for all values of } t \text{ for which } G(t) \text{ is finite.}$$

We note that:

- $0 \leq p_k \leq 1$ and $\sum_{k=0}^{\infty} p_k = 1$ by the definition of the (well-defined) probability mass function for a discrete random variable X .
- The PGF $G(t)$ is either a finite series or an infinite series.
 - ▶ If $G(t)$ is a finite series then $G(t)$ is defined for all $t \in \mathbb{R}$.
 - ▶ If $G(t)$ is an infinite series, then it is a power series and is therefore finite (or convergent) only for t in the interval of convergence of the power series.
- A PGF $G(t)$ defines a discrete random variable X (and its probability distribution) uniquely. Conversely, if X is a discrete random variable which takes values in $\mathbb{N}$, then its PGF $G(t)$ is unique.

Example 27

- a Let X be the discrete random variable which takes values 1, 2, 3, and 6, each with probability $\frac{1}{4}$. Find the PGF for X .
- b Let $X \sim B(1, \frac{1}{6})$ be the Bernoulli random variable equal to the number of '6's obtained when an unbiased 6-sided die is rolled once. Find the PGF for X .
- c Let $X \sim \text{Geo}(\frac{1}{6})$ be the geometric random variable equal to the number of rolls of an unbiased 6-sided die required to roll a '6'. Find the PGF for X .

- a $G(t) = p_1 t^1 + p_2 t^2 + p_3 t^3 + p_6 t^6 = \frac{1}{4}(t + t^2 + t^3 + t^6)$
Since $G(t)$ is a finite series, $G(t)$ is finite and therefore defined for all $t \in \mathbb{R}$.

- b $P(X = 0) = p_0 = \frac{5}{6}$
 $P(X = 1) = p_1 = \frac{1}{6}$
 $P(X = k) = 0$ for integers $k \geq 2$.
 $\therefore$ the PGF for X is $G(t) = p_0 + p_1 t + p_2 t^2 + \dots$
 $= \frac{5}{6} + \frac{1}{6}t$ for $t \in \mathbb{R}$.

- c X takes values 1, 2, 3, and $P(X = k) = \frac{1}{6} \left(1 - \frac{1}{6}\right)^{k-1}$ for $k = 1, 2, 3, \dots$
 $= \frac{1}{6} \left(\frac{5}{6}\right)^{k-1}$

$$\begin{aligned}
 \therefore \text{ the PGF for } X \text{ is } G(t) &= \sum_{k=1}^{\infty} \frac{1}{6} \left(\frac{5}{6}\right)^{k-1} t^k \\
 &= \frac{t}{6} \sum_{k=1}^{\infty} \left(\frac{5t}{6}\right)^{k-1} \\
 &= \frac{t}{6} \sum_{i=0}^{\infty} \left(\frac{5t}{6}\right)^i \quad \{\text{Infinite GS}\} \\
 &= \frac{t}{6} \times \frac{1}{1 - \frac{5t}{6}} \quad \text{if and only if } \left|\frac{5t}{6}\right| < 1 \\
 &= \frac{t}{6} \left(\frac{6}{6 - 5t}\right) \quad \text{if and only if } |t| < \frac{6}{5} \\
 &= \frac{t}{6 - 5t} \quad \text{for } t \in]-\frac{6}{5}, \frac{6}{5}[.
 \end{aligned}$$

Example 28

- a** Let X be the discrete random variable equal to the outcome of rolling an unbiased 4-sided (tetrahedral) die labelled 1, 2, 3, 4.
- i** Show that $X \sim \text{DU}(4)$. **ii** Find the PGF for X .
- b** Derive the PGF for the discrete uniform random variable $X \sim \text{DU}(n)$ which takes the values $x = 1, 2, \dots, n$.

- a i** X has probability distribution:

x	1	2	3	4
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$$\therefore X \sim \text{DU}(4)$$

$$\begin{aligned}
 \text{ii } G(t) &= p_0 + p_1t + p_2t^2 + \dots \\
 &= p_1t^1 + p_2t^2 + p_3t^3 + p_4t^4 \\
 &= \frac{1}{4}(t + t^2 + t^3 + t^4) \\
 &= \frac{t}{4}(1 + t + t^2 + t^3) \quad \{\text{Finite GS}\} \\
 &= \frac{t}{4} \frac{(t^4 - 1)}{(t - 1)}, \quad t \in \mathbb{R}
 \end{aligned}$$

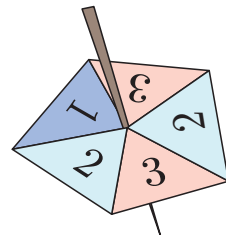
- b** Suppose $X \sim \text{DU}(n)$, $n \in \mathbb{Z}^+$.

$$\therefore P(X = x) = p_x = \frac{1}{n} \quad \text{for } x = 1, 2, 3, \dots, n.$$

$$\begin{aligned}
 \therefore G(t) &= p_0 + p_1t + p_2t^2 + \dots \\
 &= p_1t^1 + p_2t^2 + \dots + p_nt^n \\
 &= \frac{1}{n}(t + t^2 + \dots + t^n) \\
 &= \frac{t}{n}(1 + t + t^2 + \dots + t^{n-1}) \quad \{\text{Finite GS}\} \\
 &= \frac{t}{n} \frac{(t^n - 1)}{(t - 1)}, \quad t \in \mathbb{R}.
 \end{aligned}$$

EXERCISE D.1

- 1** Find the probability generating function for the discrete random variable X which takes values:
- a** 1, 2, or 3, each with probability $\frac{1}{3}$.
- b** 1, 2, or 5, each with probability $\frac{1}{3}$.
- c** 1, 2, 7, or 12 with probabilities $\frac{2}{11}$, $\frac{3}{11}$, $\frac{5}{11}$, and $\frac{1}{11}$ respectively.
- 2** Let X be the discrete random variable equal to the outcome of spinning the regular pentagonal spinner shown.
- a** Find the probability distribution of X .
- b** Find the PGF for X .



- 3** Let $X \sim B(3, \frac{2}{5})$ be a binomial random variable.
- a** Find the probability distribution of X . **b** Find the PGF for X .
- 4** Let $X \sim B(1, p)$ be a Bernoulli random variable.
- a** Show that $G(t) = 1 - p + pt$. **b** Hence find the PGF for $X \sim B(1, 0.4)$.
- 5**
- a** Find the PGF $G(t)$ for the random variable X equal to the number of heads obtained when tossing a fair coin once.
- b** Find the PGF $H(t)$ for the random variable Y equal to the number of heads obtained when tossing a fair coin twice.
- c** Verify that $H(t) = (G(t))^2$.
- 6** Let X be the random variable equal to the number of '6's obtained when an unbiased 6-sided die is rolled four times.
- a** Find the PGF $H(t)$ for $X \sim B(4, \frac{1}{6})$.
- b** Verify that $H(t) = (G(t))^4$, where $G(t) = \frac{5}{6} + \frac{1}{6}t$ is the PGF found in **Example 27**, part **b**.
- 7**
- a** Let $X \sim DU(6)$ be the random variable which has values $x = 1, 2, 3, 4, 5, 6$. Find the PGF $G(t)$ for X .
- b** Let $Y \sim DU(4)$ be the random variable which takes values $y = 1, 2, 3, 4$. Find the PGF $H(t)$ for Y .
- c** Let U be the random variable equal to the sum of values when an unbiased 6-sided die (labelled 1, 2, 3, 4, 5, 6) and a tetrahedral die (labelled 1, 2, 3, 4) are rolled.
- i** Find the probability distribution of U . **ii** Find the PGF $K(t)$ of U .
- d** Verify that $K(t) = G(t)H(t)$.
- 8** Let X be the random variable equal to the number of rolls required to roll a '4' when a tetrahedral die labelled 1, 2, 3, 4 is rolled.
- a** Show that $X \sim \text{Geo}(\frac{1}{4})$.
- b** Find:
- i** $P(X = 1) = p_1$ **ii** $P(X = 2) = p_2$ **iii** $P(X = k) = p_k$
- c** Find the PGF $G(t)$ for X in simplest form, and state the domain of $G(t)$.

PROBABILITY GENERATING FUNCTIONS FOR IMPORTANT DISTRIBUTIONS

In **Example 28** we found that for $X \sim DU(n)$, $n \in \mathbb{Z}^+$, which takes values $k = 1, 2, \dots, n$, the PGF for X is $G(t) = \frac{1}{n}(t + t^2 + \dots + t^n)$

$$= \frac{t}{n} \frac{(t^n - 1)}{(t - 1)} \quad \text{for all } t \in \mathbb{R}.$$

The PGFs for the important distributions we study in this course are summarised in the following table:

Distribution	Notation	Probability mass function $P(x)$	Probability generating function $G(t)$
Discrete uniform	$X \sim \text{DU}(n)$	$\frac{1}{n}$ for $x = 1, 2, 3, \dots, n$	$\frac{t}{n} \frac{(t^n - 1)}{t - 1}$, $t \in \mathbb{R}$
Bernoulli	$X \sim \text{B}(1, p)$, $0 < p < 1$	$p^x(1-p)^{1-x}$ for $x = 0, 1$	$1 - p + pt$, $t \in \mathbb{R}$
Poisson	$X \sim \text{Po}(m)$	$\frac{m^x e^{-m}}{x!}$ for $x = 0, 1, 2, \dots$	$e^{m(t-1)}$, $t \in \mathbb{R}$
Binomial	$X \sim \text{B}(n, p)$, $0 < p < 1$	$\binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, 3, \dots, n$	$(1 - p + pt)^n$, $t \in \mathbb{R}$
Geometric	$X \sim \text{Geo}(p)$, $0 < p < 1$	$p(1-p)^{x-1}$ for $x = 1, 2, 3, 4, \dots$	$\frac{pt}{1 - t(1-p)}$, $ t < \frac{1}{1-p}$
Negative binomial	$X \sim \text{NB}(r, p)$, $r \in \mathbb{Z}^+$, $0 < p < 1$	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	$\left(\frac{pt}{1 - t(1-p)} \right)^r$, $ t < \frac{1}{1-p}$

Example 29

Let $X \sim \text{Geo}(p)$, $0 < p < 1$ be a geometric random variable which takes values $x = 1, 2, 3, \dots$ with probabilities $P(X = x) = p(1-p)^{x-1}$.

Prove that the PGF of X is $G(t) = \frac{pt}{1 - t(1-p)}$ for $|t| < \frac{1}{1-p}$.

$$\begin{aligned}
 G(t) &= \sum_{x=1}^{\infty} P(X = x)t^x \\
 &= \sum_{x=1}^{\infty} p(1-p)^{x-1}t^x \\
 &= \frac{p}{1-p} \sum_{x=1}^{\infty} [t(1-p)]^x \quad \text{which is a GS with } u_1 = t(1-p) \\
 &\quad \text{and } r = t(1-p) \\
 &= \frac{p}{1-p} \times \frac{t(1-p)}{1 - t(1-p)} \quad \text{provided } |t(1-p)| < 1 \\
 &= \frac{pt}{1 - t(1-p)} \quad \text{provided } |t| < \frac{1}{1-p}
 \end{aligned}$$

Example 30

Let $X \sim \text{NB}(r, p)$, $r \in \mathbb{Z}^+$, $0 < p < 1$ be a negative binomial random variable which takes values $x = r, r+1, r+2, \dots$ with probabilities $P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$.

Given that $\sum_{i=0}^{\infty} \binom{r-1+i}{r-1} [t(1-p)]^i = \frac{1}{(1-t(1-p))^r}$ provided $|t(1-p)| < 1$, find the PGF for X .

The PGF for X is

$$\begin{aligned}
 G(t) &= \sum_{x=r}^{\infty} P(X=x) t^x \\
 &= \sum_{x=r}^{\infty} \binom{x-1}{r-1} p^r (1-p)^{x-r} t^x \\
 &= (pt)^r \sum_{x=r}^{\infty} \binom{x-1}{r-1} (1-p)^{x-r} t^{x-r} \\
 &= (pt)^r \sum_{i=0}^{\infty} \binom{r-1+i}{r-1} (1-p)^i t^i && \{\text{letting } i = x - r\} \\
 &= (pt)^r \sum_{i=0}^{\infty} \binom{r-1+i}{r-1} [t(1-p)]^i && \{\text{since } \binom{r-1+i}{r-1} = \binom{r-1+i}{i}\} \\
 &= (pt)^r \frac{1}{(1-t(1-p))^r} && \text{provided } |t(1-p)| < 1 \quad \{\text{given result}\} \\
 &= \left[\frac{pt}{1-t(1-p)} \right]^r && \text{provided } |t| < \frac{1}{1-p}
 \end{aligned}$$

EXERCISE D.2

- 1 Let $X \sim \text{B}(1, p)$, $0 < p < 1$, be a Bernoulli random variable. Show that the PGF for X is $G(t) = 1 - p + pt$.
- 2 Let $X \sim \text{Po}(m)$, $m > 0$ be a Poisson random variable which takes values $k = 0, 1, 2, 3, \dots$.
Given that $P(X = k) = \frac{m^k e^{-m}}{k!}$, prove that the PGF for X is $G(t) = e^{m(t-1)}$ for all $t \in \mathbb{R}$.
- 3 Let $X \sim \text{B}(n, p)$, $n \in \mathbb{Z}^+$, $0 < p < 1$, be a binomial random variable.
Given that $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, prove that the PGF of X is $G(t) = (1 - p + pt)^n$.
- 4 State the PGF for X if:
 - a $X \sim \text{Po}(6)$
 - b $X \sim \text{B}(10, 0.35)$
 - c $X \sim \text{Geo}(0.7)$
 - d $X \sim \text{NB}(6, 0.2)$
- 5 Consider a binomial random variable with mean $\mu = np$.
 - a Show that the variable has PGF $G(t) = \left(1 + \frac{\mu(t-1)}{n}\right)^n$.
 - b What happens to the PGF in a for large n ? Explain the significance of this result.

MEAN AND VARIANCE

Theorem 11

Let X be a discrete random variable with values in $\mathbb{N}$ and with PGF $G(t) = p_0 + p_1t + p_2t^2 + \dots$.

Then: **1** $G(1) = 1$

2 $E(X) = G'(1)$

3 $G''(1) = E(X(X-1)) = E(X^2) - E(X)$

4 $\text{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$

Proof:

1 $G(1) = p_0 + p_1 + p_2 + \dots = \sum p_i = 1$ by definition of the probability mass function for X .

2 $G(t) = \sum_{k=0}^{\infty} p_k t^k$

$\therefore G'(t) = \sum_{k=0}^{\infty} k p_k t^{k-1}$

$\therefore G'(1) = \sum_{k=0}^{\infty} k p_k$
 $= \sum_{k=0}^{\infty} k P(X = k)$
 $= E(X)$ by definition of $E(X)$.

3 $G''(t) = \sum_{k=0}^{\infty} k(k-1) p_k t^{k-2}$

$\therefore G''(1) = \sum_{k=0}^{\infty} k(k-1) p_k$
 $= \sum_{k=0}^{\infty} k(k-1) P(X = k)$
 $= E(X(X-1))$ {by **Theorem 2** with $g(X) = X(X-1)$ }
 $= E(X^2 - X)$
 $= E(X^2) - E(X)$ {by Corollary to **Theorem 2**}

4 $\text{Var}(X) = E(X^2) - \{E(X)\}^2$ {**Theorem 3**}
 $= E(X^2) - E(X) + E(X) - \{E(X)\}^2$
 $= G''(1) + G'(1) - [G'(1)]^2$

Example 31

Let $X \sim \text{DU}(n)$ with values $1, 2, \dots, n$. X has PGF $G(t) = \frac{1}{n}(t + t^2 + \dots + t^n)$.

Use $G(t)$ and differentiation rules to prove that $E(X) = \frac{n+1}{2}$ and $\text{Var}(X) = \frac{n^2-1}{12}$.

$$G(t) = \frac{1}{n}(t + t^2 + \dots + t^n)$$

$$G'(t) = \frac{1}{n}(1 + 2t + 3t^2 + \dots + nt^{n-1})$$

$$\begin{aligned}\therefore G'(1) &= \frac{1}{n}(1 + 2 + 3 + \dots + n) \\ &= \frac{1}{n} \left(\frac{n(n+1)}{2} \right) \\ &= \frac{n+1}{2}\end{aligned}$$

$$\therefore E(X) = G'(1) = \frac{n+1}{2}$$

$$G''(t) = \frac{1}{n}(2 + 3 \times 2t + 4 \times 3t^2 + \dots + n(n-1)t^{n-2})$$

$$\begin{aligned}\therefore G''(1) &= \frac{1}{n}(2 + 6 + 12 + \dots + n(n-1)) \\ &= \frac{1}{n} \left\{ \sum_{i=1}^{n-1} i(i+1) \right\} \\ &= \frac{1}{n} \left\{ \sum_{i=1}^{n-1} i^2 + \sum_{i=1}^{n-1} i \right\} \\ &= \frac{1}{n} \left\{ \frac{n(n-1)(2n-1)}{6} + \frac{n(n-1)}{2} \right\} \quad \{\text{well known summation identities}\} \\ &= \frac{n(n-1)}{n} \left\{ \frac{2n-1}{6} + \frac{1}{2} \right\} \\ &= (n-1) \left\{ \frac{2n+2}{6} \right\} \\ &= \frac{(n-1)(n+1)}{3}\end{aligned}$$

$$\begin{aligned}\therefore \text{Var}(X) &= G''(1) + G'(1) - \{G'(1)\}^2 \\ &= \frac{(n-1)(n+1)}{3} + \frac{n+1}{2} - \frac{(n+1)^2}{4} \\ &= \frac{(n+1)(4(n-1) + 6 - 3(n+1))}{12} \\ &= \frac{(n+1)(n-1)}{12} \\ &= \frac{n^2-1}{12}\end{aligned}$$

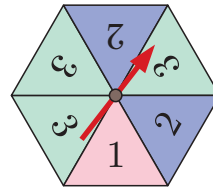
The mean and variance for some important discrete random variables are summarised below:

Distribution			$E(X)$	$\text{Var}(X)$
Discrete uniform	$X \sim \text{DU}(n)$	with values $1, 2, \dots, n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Binomial	$X \sim \text{B}(n, p)$	with values $0, 1, 2, \dots, n$	np	$np(1-p)$
Geometric	$X \sim \text{Geo}(p)$	with values $1, 2, 3, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$X \sim \text{Po}(m)$	with values $0, 1, 2, \dots$	m	m
Negative binomial	$X \sim \text{NB}(r, p)$	with values $r, r+1, r+2, \dots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

EXERCISE D.3

- 1 The PGF for random variable X , the number of heads obtained when tossing an unbiased coin, is $G(t) = \frac{1}{2} + \frac{1}{2}t$. Use $G(t)$ to find $E(X)$ and $\text{Var}(X)$.

- 2 a Write down the PGF $G(t)$ for the random variable X with value the outcome of spinning the regular hexagonal spinner shown.



- b Use $G(t)$ to find $E(X)$ and $\text{Var}(X)$.

- 3 Let X be the number of defective items leaving a production line each hour, where $X \sim \text{Po}(12)$.

- a Write down the PGF $G(t)$ of X . b Use $G(t)$ to find $E(X)$ and $\text{Var}(X)$.

- 4 Let $X \sim \text{B}(n, p)$ with values $0, 1, 2, \dots, n$. X has PGF $G(t) = (1-p+pt)^n$.

Use $G(t)$ and differentiation rules to prove that $E(X) = np$ and $\text{Var}(X) = np(1-p)$.

- 5 Let $X \sim \text{Geo}(p)$ with values $1, 2, 3, \dots$. X has PGF $G(t) = \frac{pt}{1-t(1-p)}$.

Use $G(t)$ and differentiation rules to prove that $E(X) = \frac{1}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2}$.

- 6 Let $X \sim \text{Po}(m)$ with values $0, 1, 2, \dots$. X has PGF $G(t) = e^{m(t-1)}$.

Use $G(t)$ and differentiation rules to prove that $E(X) = \text{Var}(X) = m$.

- 7 Let $X \sim \text{NB}(r, p)$ where $r \in \mathbb{Z}^+$, $0 < p < 1$. X has PGF $G(t) = \left(\frac{pt}{1-t(1-p)} \right)^r$ for $|t| < \frac{1}{1-p}$.

Use $G(t)$ and differentiation rules to prove that $E(X) = \frac{r}{p}$ and $\text{Var}(X) = \frac{r(1-p)}{p^2}$.

THE SUM OF INDEPENDENT RANDOM VARIABLES

Theorem 12

Let X and Y be two discrete random variables with values in $\mathbb{N}$ and with probability generating functions $G_X(t)$ and $G_Y(t)$ respectively.

If X and Y are **independent** then the random variable $X + Y$ has probability generating function $G_{X+Y}(t)$ where $G_{X+Y}(t) = G_X(t)G_Y(t)$.

Proof:

Suppose $G_X(t) = p_0 + p_1t + p_2t^2 + \dots$

and $G_Y(t) = q_0 + q_1t + q_2t^2 + \dots$

Let $U = X + Y$ be the random variable equal to the sum of the values of X and Y .

$$\begin{aligned} \text{Now } P(U = r) &= \sum_{k=0}^r P(X = k \text{ and } Y = r - k) \\ &= \sum_{k=0}^r P(X = k) P(Y = r - k) \quad \{\text{since } X, Y \text{ are independent}\} \\ &= \sum_{k=0}^r p_k q_{r-k} \end{aligned}$$

$$\begin{aligned} \therefore G_{X+Y}(t) &= \sum_{r=0}^{\infty} \left(\sum_{k=0}^r p_k q_{r-k} \right) t^r \\ &= p_0 q_0 + (p_0 q_1 + p_1 q_0)t + \dots + (p_0 q_r + p_1 q_{r-1} + \dots + p_r q_0)t^r + \dots \end{aligned}$$

Now consider the product

$$G_X(t)G_Y(t) = (p_0 + p_1t + p_2t^2 + \dots)(q_0 + q_1t + q_2t^2 + \dots).$$

By multiplying and collecting like terms we obtain the same function

$$p_0 q_0 + (p_0 q_1 + p_1 q_0)t + \dots + (p_0 q_r + p_1 q_{r-1} + \dots + p_r q_0)t^r + \dots = G_{X+Y}(t)$$

Corollary:

Suppose $X_1, X_2, \dots, X_n$ are independent discrete random variables with values in $\mathbb{N}$ and probability generating functions $G_{X_1}(t), G_{X_2}(t), \dots, G_{X_n}(t)$ respectively. The random variable $U = X_1 + X_2 + \dots + X_n$ has probability generating function $G_U(t) = G_{X_1}(t)G_{X_2}(t) \dots G_{X_n}(t)$.

Proof:

By the above theorem,

$$\begin{aligned} &G_{X_1+X_2+\dots+X_n}(t) \\ &= G_{X_1+\dots+X_{n-1}}(t)G_{X_n}(t) \quad \{\text{letting } X = X_1 + \dots + X_{n-1} \text{ and } Y = X_n\} \\ &= (G_{X_1+\dots+X_{n-2}}(t)G_{X_{n-1}}(t))G_{X_n}(t) \quad \{\text{letting } X = X_1 + \dots + X_{n-1} \text{ and } Y = X_{n-1}\} \\ &\vdots \\ &= G_{X_1}(t)G_{X_2}(t) \dots G_{X_n}(t) \quad \text{as required.} \end{aligned}$$

Example 32

Consider a binomial random variable $X \sim B(n, p)$ for constants $n \in \mathbb{Z}^+$, $0 < p < 1$.

Let $X = Y_1 + Y_2 + \dots + Y_n$ where $Y_1, \dots, Y_n$ are n independent Bernoulli random variables $Y_i \sim B(1, p)$, $i = 1, \dots, n$.

Find the PGF for X in terms of the PGFs for $Y_1, \dots, Y_n$.

For each Bernoulli random variable $Y_i \sim B(1, p)$,
 $G_{Y_i}(t) = 1 - p + pt$ and $X = Y_1 + Y_2 + \dots + Y_n$

$$\begin{aligned} \therefore G_X(t) &= G_{Y_1}(t)G_{Y_2}(t) \dots G_{Y_n}(t) \quad \{\text{Theorem 12}\} \\ &= (1 - p + pt)(1 - p + pt) \dots (1 - p + pt) \\ &= (1 - p + pt)^n \end{aligned}$$

Theorem 13

If X and Y are two independent discrete random variables with PGFs $G(t)$ and $H(t)$ respectively, then:

- 1 $E(X + Y) = E(X) + E(Y)$
- 2 $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

We note that:

- **Theorem 13** is only a special case of the results proved in **Theorem 6** and **Theorem 7**, which apply to any two random variables X and Y (either both discrete or both continuous).
- X and Y are required to be independent for **Theorem 13**, but **Theorem 6** shows that $E(X + Y) = E(X) + E(Y)$ holds also when X and Y are dependent.

Using the results from **Section A**, we obtain:

Corollary:

Let $X_1, X_2, \dots, X_n$ be n independent discrete random variables with PGFs $G_1(t), G_2(t), \dots, G_n(t)$. Then $E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$
 and $\text{Var}(X_1 + X_2 + \dots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)$.

Example 33

Let $X \sim \text{DU}(6)$ with values 1, 2, 3, 4, 5, 6, be the score obtained when a fair 6-sided die is rolled. Let $Y \sim \text{DU}(4)$ with values 1, 2, 3, 4, be the score obtained when a tetrahedral die is rolled.

- a Use the method of PGFs to find the probability distribution of $X + Y$.
- b Hence find $E(X + Y)$ and $\text{Var}(X + Y)$.
- c Check your answers to b using the formulae for the mean and variance of a discrete uniform distribution.

a X has PGF $G_X(t) = \frac{t}{6} \left(\frac{t^6 - 1}{t - 1} \right) = \frac{1}{6}(t + t^2 + \dots + t^6)$.

Y has PGF $G_Y(t) = \frac{t}{4} \left(\frac{t^4 - 1}{t - 1} \right) = \frac{1}{4}(t + t^2 + t^3 + t^4)$.

Since X and Y are independent random variables, $X + Y$ has PGF

$$G(t) = G_X(t)G_Y(t) = \frac{1}{6}(t + t^2 + \dots + t^6) \times \frac{1}{4}(t + t^2 + t^3 + t^4) \\ = \frac{1}{24}(t^2 + 2t^3 + 3t^4 + 4t^5 + 4t^6 + 4t^7 + 3t^8 + 2t^9 + t^{10})$$

$\therefore$ the probability distribution for $X + Y$ is:

$x + y$	2	3	4	5	6	7	8	9	10
$P(x + y)$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{24}$

b $G'(t) = \frac{1}{24}(2t + 6t^2 + 12t^3 + 20t^4 + 24t^5 + 28t^6 + 24t^7 + 18t^8 + 10t^9)$

$$\therefore G'(1) = \frac{1}{24}(2 + 6 + 12 + 20 + 24 + 28 + 24 + 18 + 10) \\ = \frac{1}{24} \times 144 \\ = 6$$

$$\therefore E(X + Y) = G'(t) = 6.$$

$$G''(t) = \frac{1}{24}(2 + 12t + 36t^2 + 80t^3 + 120t^4 + 168t^5 + 168t^6 + 144t^7 + 90t^8)$$

$$\therefore G''(1) = \frac{1}{24} \times 820 \\ \approx 34.167$$

$$\therefore \text{Var}(X + Y) = G''(1) + G'(1) - [G'(1)]^2 \\ \approx 34.167 + 6 - 6^2 \\ \approx 4.17$$

c $E(X) = \frac{6+1}{2} = 3.5$ $\text{Var}(X) = \frac{6^2 - 1}{12} = \frac{35}{12}$

$$E(Y) = \frac{4+1}{2} = 2.5$$

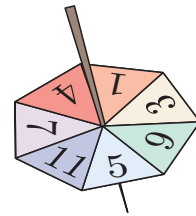
$$\text{Var}(Y) = \frac{4^2 - 1}{12} = \frac{15}{12}$$

$$\therefore E(X + Y) = E(X) + E(Y) \quad \text{and} \quad \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \\ = 3.5 + 2.5 \quad \quad \quad = \frac{35}{12} + \frac{15}{12} \\ = 6 \quad \quad \quad = 4\frac{1}{6} \\ \quad \quad \quad \approx 4.17$$

EXERCISE D.4

- 1** Let X be the random variable equal to the outcome of tossing an unbiased disc labelled 1 and 2. Let Y be the random variable equal to the outcome when rolling a tetrahedral die labelled 1, 2, 3, 4.
 - a** Find the PGF $G(t)$ for X .
 - b** Find the PGF $H(t)$ for Y .
 - c** Let $U = X + Y$ be the total score obtained when both the disc and the die are tossed and rolled.
 - i** Use $G(t)$ and $H(t)$ to write down the PGF for U in expanded form.
 - ii** Hence find $P(U = 4)$.

- 2** The random variable X equals the total score obtained when tossing a tetrahedral die labelled 1, 2, 3, 4 and a fair 6-sided die labelled 1, 2, 3, 4, 5, 6.
- Use probability generating functions to determine the probability distribution of X .
 - Hence find $P(X = 5)$.
- 3** Let $X \sim \text{Po}(m)$ and $Y \sim \text{Po}(\lambda)$ be two independent Poisson random variables.
- Use the method of PGFs to determine the PGF of $U = X + Y$.
 - Hence show that $X + Y \sim \text{Po}(m + \lambda)$.
- 4** Let $X \sim \text{B}(n, p)$ and $Y \sim \text{B}(m, p)$ be two independent binomial random variables with common probability p .
- Use the method of PGFs to determine the PGF of $U = X + Y$.
 - Hence show that $X + Y \sim \text{B}(n + m, p)$.
- 5** Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$ be two independent geometric random variables with common probability p .
- Use the method of PGFs to determine the PGF of $U = X + Y$.
 - Hence show that $X + Y \sim \text{NB}(2, p)$.
- 6** Let $X \sim \text{NB}(r, p)$ and $Y \sim \text{NB}(s, p)$ be two independent negative binomial random variables with common probability p .
- Use the method of PGFs to determine the PGF of $U = X + Y$.
 - Hence show that $X + Y \sim \text{NB}(r + s, p)$.
- 7** The spinner shown has equal chance of landing on any given side.
- State the probability of spinning a '4'.
 - Let $Y \sim \text{Geo}(\frac{1}{7})$ be the random variable equal to the number of spins required to spin a '4'.
 - Write down the PGF $G(t)$ of Y . State the domain of $G(t)$ as an interval of $\mathbb{R}$.
 - Use $G(t)$ to find $E(Y)$ and $\text{Var}(Y)$.
 - Let $X \sim \text{NB}(3, \frac{1}{7})$ be the random variable equal to the number of spins required to spin three '4's'.
 - Explain why $X = Y_1 + Y_2 + Y_3$, where $Y_i \sim \text{Geo}(\frac{1}{7})$, $i = 1, 2, 3$.
 - Use the results of **c i** and **b ii** to find $E(X)$ and $\text{Var}(X)$.
 - Which important property of Y_1, Y_2, Y_3 have you used in **c ii**?
- 8**
- Let $Y \sim \text{B}(1, p)$, $0 < p < 1$, be a Bernoulli random variable with PGF $G(t) = 1 - p + pt$. Use $G(t)$ to find:
 - $E(Y)$
 - $\text{Var}(Y)$.
 - Let $X \sim \text{B}(5, \frac{1}{6})$ be the binomial random variable equal to the number of '6's obtained in five rolls of an unbiased 6-sided die. Consider $X = Y_1 + Y_2 + \dots + Y_5$, where $Y_i \sim \text{B}(1, \frac{1}{6})$, $i = 1, \dots, 5$.
 - Write down the PGF $H(t)$ of X .
 - Use $H(t)$ to find $E(X)$ and $\text{Var}(X)$.
 - Use the results of **a** to check your answers to **b ii**. State clearly any results you use.



- 9** Let $X \sim \text{NB}(r, p)$ for $r \in \mathbb{Z}^+$ and $0 < p < 1$. Then $X = Y_1 + Y_2 + \dots + Y_r$, is the sum of r independent geometric random variables, where $Y_i \sim \text{Geo}(p)$, $i = 1, \dots, r$.

Y_1 = the number of trials to obtain the first success.

Y_2 = the number of trials after the first success to obtain the second success.

$\vdots$

Y_r = the number of trials after the $(r - 1)$ th success to obtain the r th success.

Use the PGF for a geometric random variable $G_{Y_i}(t) = \frac{pt}{1 - t(1 - p)}$ to find the PGF for X .

- 10** Let X and Y be independent discrete random variables with PGFs $G_X(t)$ and $G_Y(t)$ respectively. Use the PGF $G(t) = G_X(t)G_Y(t)$ of $X + Y$ and differentiation rules to prove:

a $E(X + Y) = E(X) + E(Y)$

b $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

- 11** The discrete random variable X takes values $0, 1, 2, 3, \dots$ and has PGF

$$G(t) = \sum_{i=0}^{\infty} p_i t^i = p_0 + p_1 t + p_2 t^2 + \dots.$$

a Find the PGF of: **i** $X + 2$ **ii** $3X$

b For constants $a, b \in \mathbb{Z}$ with $a, b > 1$:

i Show that $aX + b$ has PGF given by $H(t) = t^b G(t^a)$.

ii Use $H(t)$ and differentiation rules to prove that $E(aX + b) = aE(X) + b$ and $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

E

DISTRIBUTIONS OF THE SAMPLE MEAN AND THE CENTRAL LIMIT THEOREM

A principal application of statistics is to make **inferences** about a **population** based on observations from a sufficiently large **sample** from the population. As the sample is used to make generalisations about the whole population it is essential to employ correct sampling methods when selecting the sample.

In order to establish correct inferences about a population from a sample, we use **random sampling** where each individual in the population is equally likely to be chosen.

PARAMETERS AND STATISTICS

A **parameter** is a numerical characteristic of a *population*.

A **statistic** is a numerical characteristic of a *sample*.

A parameter or a statistic could be the mean, a percentage, the range, the standard deviation, a proportion, or many other things.

- The **mean** of a set of discrete data is its arithmetic average, and a measure of the distribution's **centre**. We let $\bar{x}$ denote the mean of a sample, and μ denote the population mean.
- The **standard deviation** of a set of data measures the deviation between the data values and the mean. It is a measure of the variability or spread of the distribution. We let s denote the standard deviation of a sample, and σ denote the population standard deviation.

When we calculate a sample statistic to estimate the population parameter, we cannot expect it to be exactly equal to the population parameter. Some measure of reliability must therefore be given, and this is generally done using a **confidence interval** which is an interval of values within which we are “confident” the population parameter lies.

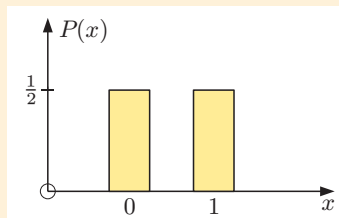
To obtain a confidence interval, we need to know how the sample statistic is distributed. We call the distribution of a sample statistic its **sampling distribution**.

CASE STUDY

COIN TOSSING

Let X be the discrete random variable equal to the number of heads obtained when an unbiased coin is tossed once. X takes values $x = 0$ (0 heads) and $x = 1$ (1 head), and $X \sim B(1, \frac{1}{2})$.

The probability distribution for X is:



TWO TOSSES

Suppose now we toss the coin twice.

Let x_1 = the number of heads on the first toss

and x_2 = the number of heads on the second toss

be the values of the random variable with distributions $X_1 \sim B(1, \frac{1}{2})$ and $X_2 \sim B(1, \frac{1}{2})$.

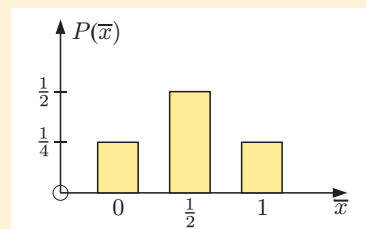
Let $\bar{x} = \frac{x_1 + x_2}{2}$ be the mean for samples of $n = 2$ tosses.

We are interested in the distribution $\bar{X}$ of $\bar{x}$, called the *sampling distribution* of $\bar{x}$, of all possible means from samples of size 2.

Possible samples	$\bar{x}$
$TT \equiv \{0, 0\}$	0
$TH \equiv \{0, 1\}$	$\frac{1}{2}$
$HT \equiv \{1, 0\}$	$\frac{1}{2}$
$HH \equiv \{1, 1\}$	1

The sampling distribution of $\bar{x}$ is:

$\bar{x}$	0	$\frac{1}{2}$	1
Frequency	1	2	1
$P(\bar{x}) = P(\bar{X} = \bar{x})$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$



THREE TOSSES

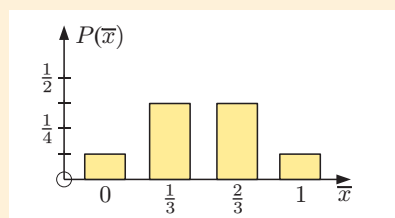
Now suppose we are interested in the sample mean, $\bar{x}$, for each possible sample of $n = 3$ tosses.

Possible samples	$\bar{x}$	Possible samples	$\bar{x}$
$TTT \equiv \{0, 0, 0\}$	0	$HHT \equiv \{1, 1, 0\}$	$\frac{2}{3}$
$TTH \equiv \{0, 0, 1\}$	$\frac{1}{3}$	$HTH \equiv \{1, 0, 1\}$	$\frac{2}{3}$
$THT \equiv \{0, 1, 0\}$	$\frac{1}{3}$	$THH \equiv \{0, 1, 1\}$	$\frac{2}{3}$
$HTT \equiv \{1, 0, 0\}$	$\frac{1}{3}$	$HHH \equiv \{1, 1, 1\}$	1

The sampling distribution $\bar{X}$ of $\bar{x}$ is:

$\bar{x}$	0	$\frac{1}{3}$	$\frac{2}{3}$	1
Frequency	1	3	3	1
$P(\bar{x}) = P(\bar{X} = \bar{x})$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

We notice that the sampling distribution $\bar{X}$ of $\bar{x}$ is again symmetric.



Example 34

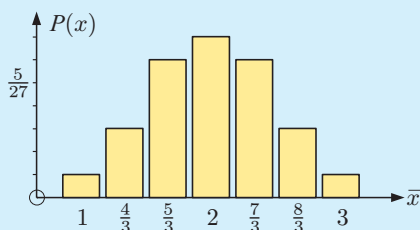
Consider an unbiased triangular spinner with possible outcomes $x = 1, 2$, or 3 . The spinner is spun three times, so $n = 3$.

- List the possible samples, and calculate the sample mean $\bar{x}$ for each.
- Draw a sampling distribution column graph to display the information.
- Describe the sampling distribution.

a	Possible samples	$\bar{x}$	Possible samples	$\bar{x}$	Possible samples	$\bar{x}$	Possible samples	$\bar{x}$
	$\{1, 1, 1\}$	1	$\{1, 3, 2\}$	2	$\{2, 2, 3\}$	$\frac{7}{3}$	$\{3, 2, 1\}$	2
	$\{1, 1, 2\}$	$\frac{4}{3}$	$\{1, 3, 3\}$	$\frac{7}{3}$	$\{2, 3, 1\}$	2	$\{3, 2, 2\}$	$\frac{7}{3}$
	$\{1, 1, 3\}$	$\frac{5}{3}$	$\{2, 1, 1\}$	$\frac{4}{3}$	$\{2, 3, 2\}$	$\frac{7}{3}$	$\{3, 2, 3\}$	$\frac{8}{3}$
	$\{1, 2, 1\}$	$\frac{4}{3}$	$\{2, 1, 2\}$	$\frac{5}{3}$	$\{2, 3, 3\}$	$\frac{8}{3}$	$\{3, 3, 1\}$	$\frac{7}{3}$
	$\{1, 2, 2\}$	$\frac{5}{3}$	$\{2, 1, 3\}$	2	$\{3, 1, 1\}$	$\frac{5}{3}$	$\{3, 3, 2\}$	$\frac{8}{3}$
	$\{1, 2, 3\}$	2	$\{2, 2, 1\}$	$\frac{5}{3}$	$\{3, 1, 2\}$	2	$\{3, 3, 3\}$	3
	$\{1, 3, 1\}$	$\frac{5}{3}$	$\{2, 2, 2\}$	2	$\{3, 1, 3\}$	$\frac{7}{3}$		

- b** The sampling distribution $\bar{X}$ of $\bar{x}$ is:

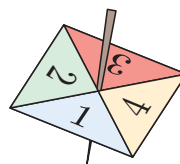
$\bar{x}$	1	$\frac{4}{3}$	$\frac{5}{3}$	2	$\frac{7}{3}$	$\frac{8}{3}$	3
Frequency	1	3	6	7	6	3	1
$P(\bar{x})$	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{6}{27}$	$\frac{7}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{1}{27}$



- c** The sampling distribution for this small value of n has a symmetric bell shape.

EXERCISE E.1

- 1** A square spinner is used to generate the digits 1, 2, 3, and 4 at random.
 - a** A sample of *two* digits is generated.
 - i** List the possible samples of $n = 2$ digits, and calculate the sample mean $\bar{x}$ for each.
 - ii** Construct a table which summarises the sampling distribution of $\bar{x}$ and the probabilities associated with it.
 - iii** Draw a sampling distribution column graph to display the information.
 - b** Repeat **a**, but this time consider samples of $n = 3$ digits.
- 2** A random variable X has two possible values (2 and 3), with equal chance of each occurring.
 - a** List all possible samples of size $n = 4$, and for each possible sample find the sample mean $\bar{x}$.
 - b** Construct a table summarising the sampling distribution of $\bar{x}$, complete with probabilities.
- 3** Two ordinary dice are rolled. The mean $\bar{x}$ of every possible set of results is calculated. Find the sampling distribution of $\bar{x}$.



ERRORS IN SAMPLING

Provided a sample is large enough, the errors should be small and the sample statistics should provide an accurate picture of the population. However, whenever sample data is collected, we expect differences between the sample statistics and population parameters.

Errors which may be due to faults in the sampling process are **systematic errors**, resulting in **bias**. Systematic errors are often due to poor sample design, or are errors made when measurements are taken.

Errors resulting from natural variability are **random errors**, sometimes called **statistical errors**.

In the following investigation we examine how well actual samples represent a population. A close look at how samples differ from each other helps us better understand the sampling error due to natural variation (random error).

INVESTIGATION 1

A COMPUTER BASED RANDOM SAMPLER

In this investigation we will examine samples from a symmetrical distribution as well as one that is skewed.

We will examine how the random process causes variations in:

- the raw data which makes up different samples
- the frequency counts of specific outcome proportions
- a measure of the centre (mean)
- a measure of the spread (standard deviation).

**What to do:**

- 1 Click on the icon. The given distribution (in column A) consists of 487 data values. The five-number summary is given and the data has been tabulated. Record the five-number summary and the frequency table given.
- 2 At the bottom of the screen click on **Samples** . Notice that the starting *sample size* is 10 and the *number of random samples* is 30. Change the *number of random samples* to 200.
- 3 Click on **Find samples** and when this is complete click on **Find sample means** .

	A	B	C	D	E	F	G	H	I	J
1	Sample Size: 10		Number of Random Samples: 30			find samples		find sample means		
2										
3	Samples:	1	2	3	4	5	6	7	8	9
4		41	46	54	46	37	14	28	78	78
5		50	61	51	58	75	68	80	90	63
6		47	79	57	48	73	58	58	72	78
7		5	44	53	66	80	69	36	19	32
8		44	62	50	28	72	41	62	39	49
9		37	27	28	59	88	36	23	53	38
10		35	15	30	31	23	32	14	77	57
11		41	50	36	64	67	69	43	54	87
12		51	88	45	76	73	29	36	54	99
13		64	50	26	50	45	72	39	45	16
14	Means:	41.50	52.20	43.00	52.60	63.30	48.80	41.90	58.10	59.7

- 4 Click on **Analyse** .
 - a Record the population mean μ , and standard deviation σ , for the population.
 - b Record the mean of the sample means and standard deviation of the sample means.
 - c Examine the associated histogram.

	A	B	C
1	Population:	Mean:	49.65
2		Standard Deviation:	19.46
3			
4	Samples:	Sample Size:	10
5		Mean of Means:	51.14
6		Standard Deviation of Means:	5.85
7	Number of Samples:		
8		30	
9			
10		Mean of Sample 1:	41.50
11		Mean of Sample 2:	52.20
12		Mean of Sample 3:	43.00

- 5 Click on **Samples** again and change the *sample size* to 20. Repeat steps 3 and 4 to gather information about the random samples of size 20.
- 6 Repeat with samples of size 30, 40, and 50. Comment on the variability.
- 7 What do you observe about the *mean of sample means* in each case, and the population mean μ ?
- 8 Is the *standard deviation of the sample means* equal to the standard deviation σ for the population?
- 9 Let the *standard deviation of the sample means* be represented by $\sigma_{\bar{X}}$. From a summary of your results, copy and complete a table like the one given. Determine the model which links $\sigma_{\bar{X}}$ and the sample size n .

n	$\sigma_{\bar{X}}^2$
10	
20	
30	
40	
50	

- 10 Now click on the icon for data from a skewed distribution. Complete an analysis of this data by repeating the above procedure. Record all your results.

STATISTICS
PACKAGE



From the **Investigation**, you should have discovered that:

- The samples consist of randomly selected members of the population.
- There is great variability in samples and their means.
- In larger samples there is less variability, so the values of $\sigma_{\bar{X}}$ are smaller.
- The means of larger samples are more accurate in estimating the population mean.
- The *mean of sample means* approximates the population mean: $\mu_{\bar{X}} \approx \mu$.
- The *standard deviation of the sample means*, $\sigma_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}}$, where n is the size of each sample.
- The distribution $\bar{X}$ of sample means $\bar{x}$, for non-normally distributed populations is approximately bell-shaped for large values of n . The larger the value of n , the more symmetric the distribution, and the more the distribution appears approximately normal.

THE SAMPLING DISTRIBUTION OF SAMPLE MEANS

Let X be any random variable with mean $\mu = E(X)$ and variance $\sigma^2 = \text{Var}(X)$.

Suppose $x_1, x_2, \dots, x_n$ are n independent observations or values taken from the distribution (or parent population) X . Each x_i has distribution X_i identical to X , and so $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$, $i = 1, \dots, n$.

Let $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$ be the *mean of a sample* $\{x_1, \dots, x_n\}$ of size n taken from the parent population X , with replacement.

By considering all possible samples of size n , the collection of possible means $\bar{x}$ forms a distribution $\bar{X}$ called the **sampling distribution of sample means**.

The distribution $\bar{X}$ has $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Proof:

$$\begin{aligned}
E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\
&= E\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n\right) \\
&= \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \dots + \frac{1}{n}E(X_n) \quad \{\text{Theorem 7}\} \\
&= \frac{1}{n}(n\mu) \\
&= \mu \\
\text{Var}(\bar{X}) &= \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\
&= \frac{1}{n^2}\text{Var}(X_1 + X_2 + \dots + X_n) \quad \{\text{Theorem 7}\} \\
&= \frac{1}{n^2}(\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)) \quad \{\text{Theorem 7}\} \\
&= \frac{1}{n^2}(n\sigma^2) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

These results confirm our observations in the previous **Investigation**.

THE CENTRAL LIMIT THEOREM (CLT)

The Central Limit Theorem gives the remarkable result that regardless of the type of distribution of the parent population X , the distribution of $\bar{X}$ of sample means will be approximately a normal distribution, provided n is large enough.

If we take samples of size n from any (normal or non-normal) population X with mean μ and variance σ^2 , then provided the sample size n is large enough, the distribution $\bar{X}$ of sample means is approximately normal and we may use the approximation $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. The larger the value of n , the better the approximation will be.

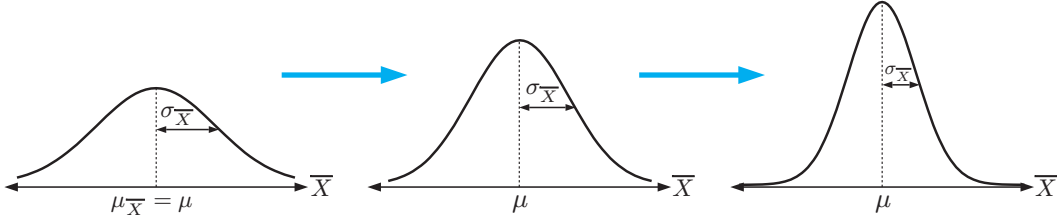
We write: $E(X) = \mu_X = \mu$ and $\text{Var}(X) = \sigma_X^2 = \sigma^2$
 $E(\bar{X}) = \mu_{\bar{X}} = \mu$ and $\text{Var}(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$.

We note that:

- Many texts suggest a “rule of thumb” of $n \geq 30$ to indicate n is large enough.
- If the parent distribution X is a normal distribution, then the size of n is not important for $\bar{X}$ to be normal. In this case, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ for all values of n , since by **Theorem 8** the mean $\bar{X}$ is $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ where $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$, and so $\bar{X}$ is a linear combination of normal random variables and is therefore itself a normal random variable.
- The distribution of $\bar{X}$ more closely approximates a normal distribution if the parent distribution of X is symmetric and not skewed.

- Since $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, the distribution of the sample means has a reducing standard deviation as n increases.
- The mean $\mu_{\bar{X}}$ is constant and equal to the population mean μ for all $n \in \mathbb{Z}^+$.

As the sample size n increases:



Example 35

A random sample of size 45 is taken from a population with mean 60 and standard deviation 4. Find the probability that the mean of the sample is less than 59.

Since the sample is of size $n = 45$, the CLT can be applied.

$$\therefore \bar{X} \sim N\left(60, \frac{16}{45}\right).$$

$$\therefore P(\bar{X} < 59) \approx 0.0468 \quad \{\mu_{\bar{X}} = 60, \sigma_{\bar{X}} = \sqrt{\frac{16}{45}}\}$$

With the Central Limit Theorem we are looking at the distributions of the *sample means* $\bar{x}$, not at the distribution of individual scores x from X .



Example 36

A large number of samples of size n are taken from $X \sim \text{Po}(2.5)$. Approximately 5% of the sample means are less than 2.025. Use the Central Limit Theorem to estimate n .

If $X \sim \text{Po}(2.5)$ then $E(X) = \mu = 2.5$ and $\text{Var}(X) = \sigma^2 = 2.5$.

By the CLT $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ approximately, and so $\bar{X} \sim N\left(2.5, \frac{2.5}{n}\right)$.

We require n such that $P(\bar{X} < 2.025) = 0.05$.

$$\therefore P\left(\frac{\bar{X} - 2.5}{\sqrt{\frac{2.5}{n}}} < \frac{2.025 - 2.5}{\sqrt{\frac{2.5}{n}}}\right) = 0.05 \quad \{\text{setting up } Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\}$$

$$\therefore \frac{2.025 - 2.5}{\sqrt{\frac{2.5}{n}}} = -1.645$$

$$\therefore n \approx 29.98$$

Since n is an integer, $n = 30$

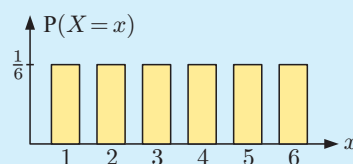
Example 37

Consider rolling a die where the random variable X is the number of dots on a face.

- a Write the probability distribution of X in a table. Graph the distribution.
- b Find the mean and standard deviation of the distribution.
- c Many hundreds of random samples of size 36 are taken. Find:
 - i $\mu_{\bar{X}}$, the mean of the sampling distribution of the sample means
 - ii $\sigma_{\bar{X}}$, the standard deviation of the sampling distribution of the sample means.
- d Comment on the shape of the distribution $\bar{X}$ of the sample means $\bar{x}$.

- a X is the discrete uniform distribution:

x_i	1	2	3	4	5	6
p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



- b $\mu = \sum p_i x_i = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \dots + \frac{1}{6}(6) = 3.5$

$$\begin{aligned}\sigma^2 &= \sum x_i^2 p_i - \mu^2 \\ &= 1\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 9\left(\frac{1}{6}\right) + 16\left(\frac{1}{6}\right) + 25\left(\frac{1}{6}\right) + 36\left(\frac{1}{6}\right) - (3.5)^2 \\ &\approx 2.9167\end{aligned}$$

$$\therefore \sigma \approx 1.708$$

- c i $\mu_{\bar{X}} = \mu = 3.5$ ii $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{36}} \approx \frac{1.708}{6} \approx 0.285 \quad \{\text{CLT}\}$

- d Since $n = 36$ is sufficiently large, we can apply the Central Limit Theorem. The distribution $\bar{X}$ of the sample means $\bar{x}$ will resemble the normal curve $\bar{X} \sim N(3.5, 0.285^2)$.

THE SAMPLING ERROR

The **sampling error** is an estimate of the margin by which the sample mean might differ from the population mean.

$\sigma_{\bar{X}}$ is used to represent the **sampling error**, also called the **standard error**, of the mean. For samples of n independent values taken from a distribution with standard deviation σ , the sampling error is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

In summary, there are two factors which help us to decide if a sample provides useful and accurate information:

- **Sample size**

If the sample size is too small, the statistics obtained from it may be unreliable. A sufficiently large sample should reflect the same mean as the population it comes from.

- **Sample error**

The sampling error indicates that for a large population, a large sample may be unnecessary. For example, the reliability of the statistics obtained from a sample of size 1000 can be almost as good as those obtained from a sample of size 4000. The additional data may provide only slightly more reliable statistics, and therefore be considered unnecessary.

EXERCISE E.2

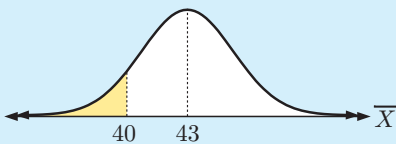
- 1 Random samples of size 36 are selected from a population with mean 64 and standard deviation 10. For the sampling distribution of the sample means, find:
- the mean
 - the standard deviation.
- 2 Random samples of size n are selected from a population with standard deviation 24.
- Write $\sigma_{\bar{X}}$ in terms of n .
 - Find $\sigma_{\bar{X}}$ when:
 - $n = 4$
 - $n = 16$
 - $n = 64$
 - How large must a sample be for the sampling error to equal 4?
 - Graph $\sigma_{\bar{X}}$ against n .
 - Discuss $\sigma_{\bar{X}}$ as n increases in value. Explain the significance of this result.
- 3 A large number of samples of size n are taken from $X \sim \text{Po}(6)$. Approximately 9% of the sample means are less than 5. Find the value of n .
- 4 The IQ measurements of a population have mean 100 and standard deviation of 15. Many hundreds of random samples of size 36 are taken from the population, and a relative frequency histogram of the sample means is formed.
- What would we expect the mean of the sample means to be?
 - What would we expect the standard deviation of the sample means to be?
 - What would we expect the shape of the histogram to look like?

Example 38

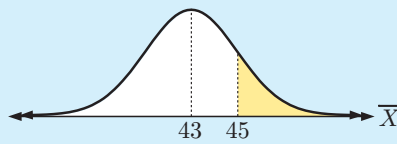
The age of business men in Sweden is normally distributed with mean 43 and standard deviation 8. If 16 business men are randomly selected from the population, find the probability that the sample mean of these measurements is:

- less than 40
- greater than 45
- between 37 and 47.

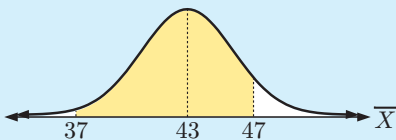
$$\bar{X} \sim N\left(43, \left(\frac{8}{\sqrt{16}}\right)^2\right), \text{ so } \bar{X} \sim N(43, 2^2).$$

a

$$P(\bar{X} < 40) \approx 0.0668$$

b

$$P(\bar{X} > 45) \approx 0.159$$

c

$$P(37 < \bar{X} < 47) \approx 0.976$$

- 5 When a coin is tossed, the random variable X is the number of heads which appear.

The probability distribution for x is:

x_i	0	1
p_i	$\frac{1}{2}$	$\frac{1}{2}$

- Find μ and σ for the X -distribution.
- Now consider the sampling distribution of $\bar{X}$.
 - List the 16 possible samples of size $n = 4$ and construct a probability function table.
 - For this sampling distribution of means in **b**, find $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$.
 - Check with **a** that $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

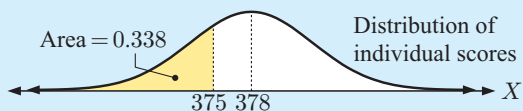
Example 39

The contents of soft drink cans is randomly distributed with mean 378 mL and standard deviation 7.2 mL. Find the likelihood that:

- an individual can contains less than 375 mL
- a box of 36 cans has average contents less than 375 mL.

a $X \sim N(378, 7.2^2)$

$P(X < 375) \approx 0.338$

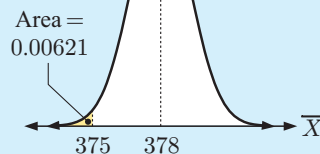


b $\bar{X} \sim N(378, \frac{7.2^2}{36})$

$P(\bar{X} < 375) \approx 0.00621$

There is approximately 0.6% chance of getting a box of 36 cans with average contents less than 375 mL, compared with a 33.8% chance of an individual can having contents less than 375 mL.

Distribution of sample means

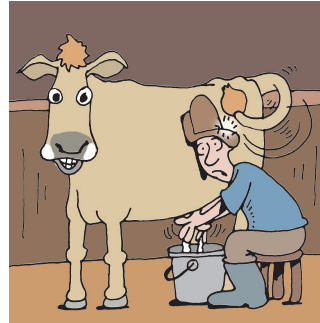


The scores for *individual cans* have distribution $X \sim N(378, 7.2^2)$.
The scores for the *means of samples of size 36* have distribution $\bar{X} \sim N(378, \frac{7.2^2}{36})$.



- The values of homes in a wealthy suburb of a small city are skewed high with mean €620 000 and standard deviation €80 000. A sample of 25 homes was taken and the mean of the sample was €643 000.
 - Using the Central Limit Theorem, find the probability that a random sample of 25 homes in this suburb will have a mean of at least €643 000.
 - Comment on the reliability of your answer to **a**.
- The heights of a particular species of plant have mean 21 cm and standard deviation $\sqrt{90}$ cm. A random sample of 40 plants is taken and the mean height is calculated. Find the probability that this sample mean lies between 19.5 cm and 24 cm.
- At a college, the masses of the male students have mean 70 kg and standard deviation 5 kg. 64 male students are chosen at random. Find the probability that their mean mass is less than 68.75 kg.

- 9 A random sample of size 100 is taken from $X \sim B(20, 0.6)$ and the sample mean $\bar{x}$ is calculated. Use the Central Limit Theorem to find the probability that:
- $\bar{x}$ is greater than 12.4
 - $\bar{x}$ is less than 12.2.
- 10 Suppose the duration of human pregnancies can be modelled by a normal distribution with mean 267 days and standard deviation 15 days. If a pregnancy lasts longer than 267 days it is said to be “overdue”. If a pregnancy lasts less than 267 days it is said to be “premature”.
- What percentage of pregnancies will be overdue by between 1 and 2 weeks?
 - Find the 80th percentile for pregnancy duration.
 - A certain obstetrician is providing prenatal care for 64 pregnant women.
 - Describe the sampling distribution $\bar{X}$ for the sample mean of all random samples of size 64.
 - Find the mean and standard deviation for the distribution of the random variable $\bar{X}$.
 - Find the probability that the mean duration of the obstetrician’s patients’ pregnancies will be premature by at least one week.
 - Suppose the duration of human pregnancies is not actually a normal distribution, but is skewed to the left. Does that change the answers to parts **a** to **d** above?
- 11 On average, Ayrshire cows produce 49 units of milk per day with standard deviation 5.87 units, whereas Jersey cows produce 44.8 units of milk per day with standard deviation 5.12 units. For each breed, the production of milk can be modelled by a normal distribution.
- Find the probability that a randomly selected Ayrshire cow will produce more than 50 units of milk daily.
 - Find the probability that a randomly selected Jersey cow will produce more milk than a randomly selected Ayrshire cow.
 - A dairy farmer has 25 Jersey cows. Find the probability that the daily production for this small herd exceeds 46 units per cow per day.
 - A neighbouring farmer has 15 Ayrshire cows. Find the probability that this herd produces at least 4 units per cow per day more than the Jersey herd.



Example 40

The weights of male employees in a bank are normally distributed with a mean $\mu = 71.5$ kg and standard deviation $\sigma = 7.3$ kg. The bank has an elevator with a maximum load of 444 kg. Six male employees enter the elevator. Calculate the probability that their combined weight exceeds the maximum load.

$$X \sim N(71.5, 7.3^2)$$

Let $\bar{X}$ be the mean weight of a random sample of $n = 6$ male employees.

$$\text{By the CLT, } \bar{X} \sim N\left(71.5, \frac{7.3^2}{6}\right)$$

$$\therefore P(\bar{X} > \frac{444}{6}) \approx 0.201 \quad \{\sigma_{\bar{X}} = \frac{7.3}{\sqrt{6}}\}$$

This is the same answer as in **Example 6 a**.

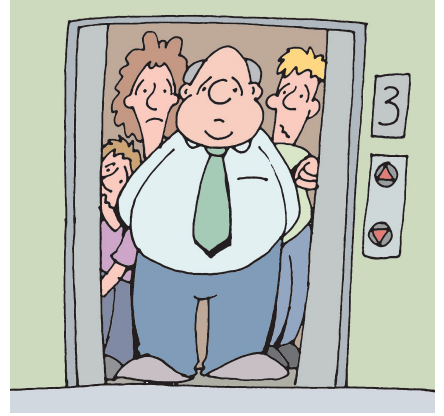
We repeat **Example 6 a**, but this time employ the Central Limit Theorem.



- 12** An elevator has a maximum recommended load of 650 kg. The weights of adult males are distributed normally with a mean of 73.5 kg and standard deviation of 8.24 kg. What is the maximum recommended number of adult males that should use the elevator at any one time, if you want to be at least 99.5% certain that the total weight does not exceed the maximum recommended load.

Hint: Start with $n = 9$.

- 13** A large number of samples of size n are taken from a population with mean 74 and standard deviation 6. The probability that the sample mean is less than 70.4, is 0.00135. Use the Central Limit Theorem to find n .



Example 41

A population is known to have a standard deviation of 8 but has an unknown mean μ . In order to estimate the mean μ , the mean of a random sample of size 60 is taken. Find the probability that the estimate is in error by less than 2.

Since $n = 60$, the CLT applies.

Let $\bar{X}$ be the mean of a random sample of size 60.

By the CLT, $\bar{X} \sim N(\mu, \frac{8^2}{60})$.

The error may be positive or negative, so it is $\bar{X} - \mu$ or $\mu - \bar{X}$, and we need to find $P(|\bar{X} - \mu| < 2)$.

$$\begin{aligned}
 \text{Now } P(|\bar{X} - \mu| < 2) &= P(-2 < \bar{X} - \mu < 2) \\
 &= P\left(\frac{-2}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{2}{\frac{\sigma}{\sqrt{n}}}\right) \quad \left\{\text{setting up } Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right\} \\
 &= P\left(\frac{-2}{\frac{8}{\sqrt{60}}} < Z < \frac{2}{\frac{8}{\sqrt{60}}}\right) \\
 &= P\left(-\frac{\sqrt{60}}{4} < Z < \frac{\sqrt{60}}{4}\right) \\
 &\approx 0.947
 \end{aligned}$$

$\therefore$ there is about 94.7% probability that the estimate is in error by less than 2.

- 14** The standard deviation of the masses of articles in a large population is 4.55 kg. If random samples of size 100 are drawn from the population, find the probability that a sample mean will differ from the true population mean by less than 0.8 kg.
- 15** The lengths of rods produced by a machine have mean 100 cm and standard deviation 15 cm. Find the probability that if 60 rods are randomly chosen from the machine, the mean length of the sample will be at least 105 cm.

- 16** Chocblock bite-size chocolate bars are produced in a factory using a machine. The weights of the bars are normally distributed with mean 18.2 grams and standard deviation 3.3 grams. The bars are sold in packets containing 25 bars each. Hundreds of thousands of packets are produced each year.
- Find $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ for this situation.
 - On each packet it is printed that the nett weight of contents is 425 grams.
 - What is the manufacturer claiming about the mean weight of each bar?
 - What percentage of packets will fail to meet the 425 gram claim?
 - Suppose an additional bar is added to each packet with the nett weight claim retained at 425 grams.
 - Find the new values of $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$.
 - What percentage of packets will fail to meet the claim now?

- 17** In **Example 8**, a cereal manufacturer produces packets of cereal in two sizes, small (S) and economy (E). The amount in each packet is distributed normally and independently as shown in the table.

	Mean (g)	Variance (g^2)
Small	315	4
Economy	950	25

For bulk shipping, 15 small packets of cereal are placed in small cartons, and separately, 10 economy packets of cereal are placed in economy cartons.

Let $\bar{s}$ be the mean weight of a packet in a carton of small packets of cereal.

Let $\bar{e}$ be the mean weight of a packet in a carton of economy packets of cereal.

- Write down the distribution:
 - $\bar{S}$ of $\bar{s}$
 - $\bar{E}$ of $\bar{e}$.
 - A carton of each type is selected at random. Find the probability that the mean weight in the economy carton is less than three times the mean weight in the small carton.
 - One economy carton and three small cartons are selected at random. Find the probability that the mean weight in the economy carton is less than the sum of the mean weights in the small cartons.
- 18** A pharmaceutical company claims that a new drug cures 75% of patients suffering from a certain disease. However, a medical committee believes that less than 75% are cured. To test the pharmaceutical company's claim, a trial is carried out in which 100 patients suffering from the disease are given the new drug. It is found that 68 of these patients are cured.
- Assuming the pharmaceutical company's claim is true:
 - state the distribution which determines the number of patients cured in a random sample of 100 patients
 - write down the mean and the standard deviation of this distribution
 - determine the probability that 68 or fewer patients are cured in a random sample of 100 patients
 - use the CLT to determine the probability that 68 or fewer patients are cured in a random sample of 100 patients.
 - Discuss whether the company's claim is reasonable.

THE PROPORTION OF SUCCESSES IN A LARGE SAMPLE

We are frequently presented by the media with estimates of population proportions, often in the form of percentages.

For example:

- If an election was held tomorrow, 52% of the population would vote for the “Do Good” party.
- 17% of the South African population tested positive to HIV.
- 73% of company executives say they will not employ smokers.

To estimate a **population proportion** p , we consider taking a random sample. The distribution of the random variable $\hat{p}$, the **sample proportion**, determines the accuracy of the estimate.

Consider the election example:

To estimate the proportion of voters who intend to vote for the “Do Good” party, a random sample of 3500 voters was taken and 1820 indicated they would vote “Do Good”. The **sample proportion** of “Do Good” voters is denoted $\hat{p} = \frac{1820}{3500} = 0.52$.

However, this is just one particular sample. We want to know, for *all* samples of size n , the mean $\mu_{\hat{p}}$ and standard deviation $\sigma_{\hat{p}}$ of the $\hat{p}$ distribution.

Firstly, we see that $\hat{p} = \frac{X}{n}$ where $\begin{cases} \hat{p} = \text{the sample proportion} \\ X = \text{the number of successes in the sample} \\ n = \text{sample size.} \end{cases}$

The random variable X which stands for the number of successes in the sample (the number who vote “Do Good” in our example) has a binomial distribution $X \sim B(n, p)$, where p is the population proportion of “Do Good” voters.

Now consider $\hat{p} = \frac{Y_1 + Y_2 + \dots + Y_n}{n}$, where $Y_1, Y_2, \dots, Y_n$ are the n independent Bernoulli random variables $Y_i \sim B(1, p)$, with $E(Y_i) = p$, $\text{Var}(Y_i) = p(1 - p)$, $i = 1, 2, \dots, n$, such that $X = Y_1 + Y_2 + \dots + Y_n$, and

$$Y_i = \begin{cases} 0 & \text{if that person in the sample does not vote for the “Do Good” party.} \\ 1 & \text{if that person in the sample does vote for the “Do Good” party.} \end{cases}$$

Therefore $\hat{p} = \bar{Y}$ = the sample mean of sample $\{Y_1, Y_2, \dots, Y_n\}$.

If n is sufficiently large:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

$\hat{p}$ is a mean.



Proof:

$$\begin{aligned} E(\hat{p}) &= E(\bar{Y}) = E\left(\frac{Y_1 + Y_2 + \dots + Y_n}{n}\right) \\ &= \frac{1}{n}(E(Y_1) + E(Y_2) + \dots + E(Y_n)) \\ &= \frac{1}{n} \times np \quad \{\text{Theorem 6}\} \\ &= p \end{aligned}$$

$$\begin{aligned}
 \text{Var}(\hat{p}) &= \text{Var}(\bar{Y}) = \text{Var}\left(\frac{Y_1 + Y_2 + \dots + Y_n}{n}\right) \\
 &= \frac{1}{n^2}(\text{Var}(Y_1) + \text{Var}(Y_2) + \dots + \text{Var}(Y_n)) \\
 &= \frac{1}{n^2}(np(1-p)) \\
 &= \frac{p(1-p)}{n}
 \end{aligned}$$

Since $\hat{p} = \bar{Y}$ is a sample mean, and since n is large, we can apply the CLT. $\hat{p}$ has an approximately normal distribution, and $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$.

Also, since $X = n\hat{p}$, and for large n , $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$, by **Theorem 8** we find that X has an approximately normal distribution with

$$\begin{aligned}
 E(X) &= E(n\hat{p}) & \text{and} & & \text{Var}(X) &= \text{Var}(n\hat{p}) \\
 &= nE(\hat{p}) & & & &= n^2\text{Var}(\hat{p}) \\
 &= np & & & &= n^2 \times \frac{p(1-p)}{n} \\
 & & & & &= np(1-p).
 \end{aligned}$$

Therefore, the discrete binomial random variable $X \sim B(n, p)$ can be approximated by a continuous normal random variable $X_c \sim N(np, np(1-p))$.

From **Section C**, the accepted conditions to apply this normal approximation to the binomial distribution are $np > 5$ and $n(1-p) > 5$. We also now add $n \geq 30$ for the CLT to apply.

Example 42

The local paper claims that Ms Claire Burford gained 43% of the votes in the local Council elections.

- a** Find the probability that a poll of randomly selected voters would show over 50% in favour of Ms Burford, given a sample size of: **i** 150 **ii** 750
- b** A sample of 100 voters was taken and 62% of these voted for Ms Burford. Find the probability of this occurring and comment on the result.

- a i** The population proportion $p = 0.43$, so $1 - p = 0.57$.
The sample size $n = 150$ is large enough to apply the CLT.

$$\text{Now } \hat{p} \sim N\left(0.43, \frac{0.43 \times 0.57}{150}\right)$$

$$\therefore P(\hat{p} > 0.5) \approx 0.0417$$

$$\text{ii } \hat{p} \sim N\left(0.43, \frac{0.43 \times 0.57}{750}\right)$$

$$\therefore P(\hat{p} > 0.5) \approx 0.000\,054\,0$$

$$\text{b } \hat{p} \sim N\left(0.43, \frac{0.43 \times 0.57}{100}\right)$$

$$\therefore P(\hat{p} \geq 0.62) \approx 0.000\,062\,1$$

This is so unlikely that we would question the accuracy of the claim that Ms Burford only gained 43% of the vote.

EXERCISE E.3

- 1 An egg producer claims that less than 4% of the eggs delivered to a supermarket will be broken. On a busy day, 1000 eggs are delivered to this supermarket and 7% are broken.
 - a Find the probability that this will happen.
 - b Briefly comment on the producer's claim.
- 2 Two sevenths of households in a country town are known to own computers. Find the probability that in a random sample of 100 households, no more than 29 households own a computer.
- 3 A pre-election poll is run to determine the proportion of voters who favour the Labour Party (LP). The poll is based on one random sample of 2500 voters. Let p be the true proportion of voters who support the LP, and suppose $p = 0.465$.
 - a For the sampling distribution of proportions, find:
 - i the mean
 - ii the standard deviation.
 - b State the sampling distribution and its normal approximation.
 - c Hence, find the probability that:
 - i the sample mean is less than 0.46
 - ii between 45% and 47% of voters in the sample favour the LP
 - iii the proportion $\hat{p}$ of LP supporters in the sample differs by more than 0.035 from p .
 - d Interpret your answer in c iii.
- 4 Eighty five percent of the plum trees grown in a particular area produce more than 700 plums.
 - a State the sampling distribution for the proportion of plum trees that produce more than 700 plums in this area, assuming a sample of size n .
 - b State the conditions under which the sampling distribution can be approximated by the normal distribution.
 - c If a random sample of 200 plum trees is selected, find the probability that:
 - i less than 75% produce more than 700 plums
 - ii between 75% and 87% produce more than 700 plums.
 - d In a random sample of 500 plum trees, 350 produced more than 700 plums.
 - i Find the probability of 350 or fewer trees producing more than 700 plums.
 - ii Comment, giving two reasons, why this sample is possible.
- 5 A regular pentagon has sectors numbered 1, 1, 2, 3, 4. Find the probability that, when the pentagon is spun 400 times, the result of a 1 occurs:
 - a more than 150 times
 - b less than 175 times.
- 6 A tyre company in Moscow claims that at least 90% of the tyres they sell will last at least 30 000 km. A sample of 250 tyres were tested, and it was found that 200 of the tyres did not last for at least 30 000 km.
 - a State the distribution of the sample proportions, with any assumptions made.
 - b Find the proportion of samples of 250 tyres that would have no more than 200 tyres lasting at least 30 000 km.
 - c Comment on this result.



F

POINT ESTIMATION (UNBIASED ESTIMATORS AND ESTIMATES)

In the previous section we considered taking a sample of n independent observations (or values) from a random variable X with mean $E(X) = \mu$ and variance $\text{Var}(X) = \sigma^2$.

Let X be a random variable with mean μ and variance σ^2 . A sample of size n is taken from the population of X with replacement so the values are independent. The independent values $x_1, x_2, \dots, x_n$ can be interpreted as $x_1 \in X_1, x_2 \in X_2, \dots, x_n \in X_n$ where $X_i, i = 1, 2, \dots, n$ are n copies of the distribution of X .

Thus $E(X_1) = E(X_2) = \dots = E(X_n) = \mu$

and $\text{Var}(X_1) = \text{Var}(X_2) = \dots = \text{Var}(X_n) = \sigma^2$.

It is often impractical to work with an entire population, and often the population parameters μ, σ^2 , and proportion p are unknown anyway. In these cases, we work with a sample, and we use the sample statistics to estimate the population parameters.

For a large enough sample, we expect:

- the sample mean $\bar{x}$ to be close in value to the population mean μ
- the sample proportion $\hat{p}$ to be close in value to the population proportion p .

An **estimator** T is a statistic, which is a function of the values in a sample, used to estimate a population parameter θ .

An **estimate** t is a specific value of T calculated from a particular sample.

For example, the number of heads obtained when an unbiased coin is tossed once, is described by $X \sim B(1, \frac{1}{2})$, where $E(X) = \mu = \frac{1}{2}$ and $\text{Var}(X) = \sigma^2 = \frac{1}{4}$.

Suppose the coin is tossed many times. Consider samples $\{x_1, x_2, x_3\}$ of size $n = 3$, which are the numbers of heads obtained in each of three (independent) coin tosses.

The associated sample mean $\bar{x} = \frac{x_1 + x_2 + x_3}{3}$ has distribution

$$\bar{X} = \frac{X_1 + X_2 + X_3}{3}$$

= the average number of heads obtained in three coin tosses

where X_1, X_2, X_3 are identical copies of the distribution X . $\bar{X}$ is an *estimator* for the mean μ of X .

For the sample $\{1, 0, 1\}$, corresponding to the coin toss results HTH, then $\bar{x} = \frac{1+0+1}{3} = \frac{2}{3}$ is an *estimate* of μ .

Let $\hat{p}$ be the sample proportion of heads obtained in n coin tosses.

Then $\hat{p} = \frac{X_1 + X_2 + \dots + X_n}{n}$ where $X_1, X_2, \dots, X_n$ are identical copies of the distribution X .

$\hat{p}$ is an *estimator* for the population proportion $p = \frac{1}{2}$ of heads obtained when an unbiased coin is tossed many, many times.

$\hat{p}$ is also a mean!



For the sample $\{1, 1, 0, 0, 1, 0, 1\}$, corresponding to the coin toss results HHTTHTH, when the coin is tossed seven times, $\hat{p} = \frac{1+1+0+0+1+0+1}{7} = \frac{4}{7}$ is an *estimate* of p .

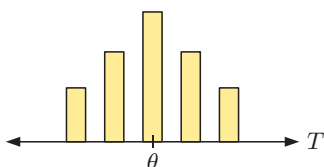
Intuitively, a “good” estimator should have values which centre about the parameter it is approximating, and not be biased to values below or above the value of the parameter.

The set of all estimates t , calculated from each possible sample using estimator T , is the **sampling distribution** of T .

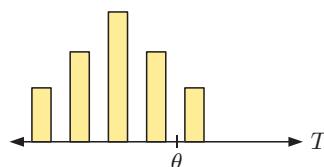
An estimator T for a population parameter θ is **unbiased** if the mean of the sampling distribution of T equals the parameter θ , so $E(T) = \theta$. Any estimate t from T is then called an **unbiased estimate** of θ .

Otherwise, T is a **biased** estimator of θ , and any estimate t from T is then called a **biased estimate** of θ .

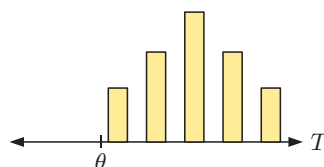
Unbiased estimator T of θ



Biased estimator T of θ



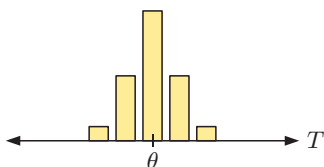
or



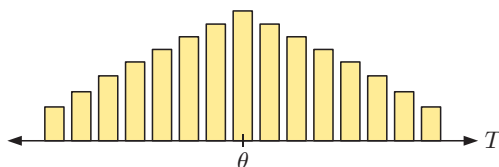
We have already proved in **Section E** that the sampling distribution of $\bar{X}$ has $E(\bar{X}) = \mu$ and the sampling distribution of $\hat{p}$ has $E(\hat{p}) = p$. Hence:

- A sample mean $\bar{x}$ is an unbiased estimate of the population mean μ .
- A sample proportion $\hat{p}$ is an unbiased estimate of the population proportion p .

Suppose we have an unbiased estimator T for a population parameter θ , so $E(T) = \theta$, and therefore estimates found using T are centred about θ . The sampling distribution of T could be:



or



We see that the *spread* of the distribution of estimates could be small or large. It follows that the estimates found using T could vary but be close to θ , or could vary wildly from the value of θ .

We therefore define:

If T_1 and T_2 are two unbiased estimators for a parameter θ , then T_1 is a **more efficient** estimator than T_2 if $\text{Var}(T_1) < \text{Var}(T_2)$.

Example 43

Let X be a random variable with unknown mean μ and unknown variance σ^2 . Consider samples $\{x_1, x_2\}$ of size 2 of independent values taken from X .

Let $T_1 = \frac{X_1 + X_2}{2}$ and $T_2 = \frac{3X_1 + 5X_2}{8}$.

- Show that T_1 is an unbiased estimator of μ .
- Show that T_2 is an unbiased estimator of μ .
- Calculate estimates t_1 and t_2 for the sample $\{2.1, 3.5\}$.
- Find $\text{Var}(T_1)$ and $\text{Var}(T_2)$.
- Which of T_1 and T_2 is the more efficient estimator of μ ? Why?

Since X_1 and X_2 each have distribution identical to the distribution of X ,
 $E(X_1) = E(X_2) = E(X) = \mu$ and $\text{Var}(X_1) = \text{Var}(X_2) = \text{Var}(X) = \sigma^2$.

$$\begin{aligned} \text{a } E(T_1) &= E\left(\frac{X_1 + X_2}{2}\right) \\ &= \frac{1}{2}E(X_1 + X_2) \\ &= \frac{1}{2}(E(X_1) + E(X_2)) \\ &= \frac{1}{2}(\mu + \mu) \\ &= \frac{2\mu}{2} \\ &= \mu \end{aligned}$$

$\therefore T_1$ is an unbiased estimator of μ .

$$\text{c } t_1 = \frac{2.1 + 3.5}{2} = \frac{5.6}{2} = 2.8$$

$$\begin{aligned} \text{d } \text{Var}(T_1) &= \text{Var}\left(\frac{1}{2}X_1 + \frac{1}{2}X_2\right) \\ &= \frac{1}{4}\text{Var}(X_1) + \frac{1}{4}\text{Var}(X_2) \\ &= \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \\ &= \frac{2\sigma^2}{4} \\ &= \frac{\sigma^2}{2} \end{aligned}$$

$$\text{e } \sigma^2 \text{ is a constant, and } \frac{\sigma^2}{2} = \frac{16}{32}\sigma^2$$

$$\text{Now, } \frac{16}{32}\sigma^2 < \frac{17}{32}\sigma^2$$

$$\therefore \text{Var}(T_1) < \text{Var}(T_2)$$

$\therefore T_1$ is the more efficient estimator of μ .

$$\begin{aligned} \text{b } E(T_2) &= E\left(\frac{3X_1}{8} + \frac{5X_2}{8}\right) \\ &= \frac{3}{8}E(X_1) + \frac{5}{8}E(X_2) \\ &= \frac{3}{8}\mu + \frac{5}{8}\mu \\ &= \mu \\ \therefore T_2 &\text{ is an unbiased estimator of } \mu. \end{aligned}$$

$$t_2 = \frac{3(2.1) + 5(3.5)}{8} = 2.975$$

$$\begin{aligned} \text{Var}(T_2) &= \text{Var}\left(\frac{3}{8}X_1 + \frac{5}{8}X_2\right) \\ &= \left(\frac{3}{8}\right)^2 \text{Var}(X_1) + \left(\frac{5}{8}\right)^2 \text{Var}(X_2) \\ &= \frac{9}{64}\sigma^2 + \frac{25}{64}\sigma^2 \\ &= \frac{34}{64}\sigma^2 \\ &= \frac{17}{32}\sigma^2 \end{aligned}$$

ESTIMATION OF VARIANCE

Consider the distribution of a random variable X with unknown mean $E(X) = \mu$ and unknown variance $\text{Var}(X) = \sigma^2$.

Let $\{x_1, x_2, \dots, x_n\}$ be a random sample of size n of independent values taken from random variable X . We are sampling with replacement, and each value x_i is a variable where $x_i \in X_i$, $i = 1, \dots, n$, and the distribution of X_i is identical to the distribution of X . Thus $E(X_i) = E(X) = \mu$ and $\text{Var}(X_i) = \text{Var}(X) = \sigma^2$ for $i = 1, \dots, n$.

$\bar{x} = \frac{x_1 + \dots + x_n}{n}$ is the **sample mean** as usual, and $s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ is the **sample variance**.

As shown above, $\bar{x}$ lies in the sampling distribution $\bar{X}$, and is an unbiased estimate of μ .

The sample variance s_n^2 lies in the sampling distribution S_n^2 of sample variances. However, we do not yet know whether s_n^2 is an unbiased or biased estimate of the population variance σ^2 .

$$\begin{aligned}
 \text{Consider } S_n^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \\
 &= \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
 &= \frac{1}{n} \left\{ \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + \sum_{i=1}^n \bar{X}^2 \right\} \\
 &= \frac{1}{n} \left\{ \sum_{i=1}^n X_i^2 - 2\bar{X}(n\bar{X}) + n\bar{X}^2 \right\} \\
 &= \frac{1}{n} \left\{ \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right\} \\
 \therefore E(S_n^2) &= E\left(\frac{1}{n} \left\{ \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right\}\right) \\
 &= \frac{1}{n} E\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) \\
 &= \frac{1}{n} \left\{ \sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right\} \\
 &= \frac{1}{n} \left\{ \sum_{i=1}^n (\text{Var}(X_i) + \{E(X_i)\}^2) - n(\text{Var}(\bar{X}) + \{E(\bar{X})\}^2) \right\} \\
 &\quad \{\text{Var}(X_i) = E(X_i^2) - \{E(X_i)\}^2 \text{ from Theorem 3}\} \\
 &= \frac{1}{n} \left\{ \sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \right\} \quad \begin{aligned} &\{\text{Var}(X_i) = \sigma^2, E(X_i) = \mu \\ &\text{Var}(\bar{X}) = \frac{\sigma^2}{n}, E(\bar{X}) = \mu\} \end{aligned} \\
 &= \frac{1}{n} \{n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\} \\
 &= \left(\frac{n-1}{n}\right) \sigma^2 \\
 &\neq \sigma^2
 \end{aligned}$$

Hence S_n^2 is a biased estimator of σ^2 , and the sample variance s_n^2 is a biased estimate of σ^2 .

For this reason, we define $S_{n-1}^2 = \left(\frac{n}{n-1}\right) S_n^2$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

Then $E(S_{n-1}^2) = E\left(\left(\frac{n}{n-1}\right) S_n^2\right)$

$$= \left(\frac{n}{n-1}\right) E(S_n^2) \quad \{\text{Theorem 7}\}$$

$$= \frac{n}{(n-1)} \times \frac{(n-1)}{n} \times \sigma^2$$

$$= \sigma^2$$

$\therefore S_{n-1}^2$ is an unbiased estimator of σ^2 .

Suppose $\{x_1, x_2, \dots, x_n\}$ is a sample of size n of independent values taken from a population X with mean μ and variance σ^2 . $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$ is the sample mean.

• $S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ is a **biased estimator** of σ^2

and therefore $s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ is a **biased estimate** of σ^2 .

• $S_{n-1}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ is an **unbiased estimator** of σ^2

and therefore $s_{n-1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ is an **unbiased estimate** of σ^2 .

We note that:

- The values $x_1, x_2, \dots, x_n$ are dependent, since $x_1 + x_2 + \dots + x_n = n\bar{x}$. Thus given any $n-1$ of these values, the n th value is completely determined. The calculation s_{n-1}^2 is therefore based on $n-1$ independent values, and we say the number of **degrees of freedom** of s_{n-1}^2 is $n-1$.
- $s_n^2 \leq s_{n-1}^2$ for all $n \in \mathbb{Z}^+$, $n \geq 2$.

Example 44

Suppose X is a random variable with unknown mean μ and unknown variance σ^2 .

Suppose $\{1.10, 2.5, 1.75, 3.45, 8.41, 6.75, 4.53\}$ is a sample of independent values taken from X .

- Calculate an unbiased estimate of μ .
- Calculate: i s_6^2 ii s_7^2
- Write down $E(S_6^2)$ and $E(S_7^2)$ in terms of σ^2 .
- Of the estimates of σ^2 found in **b**, which is the preferred estimate? Why?

- a** The sample has size $n = 7$.

$$\therefore \bar{x} = \frac{1.10 + 2.5 + 1.75 + 3.45 + 8.41 + 6.75 + 4.53}{7} = 4.07 \text{ is an unbiased estimate of } \mu.$$

$$\begin{aligned} \text{b i } s_{n-1}^2 &= s_6^2 \\ &= \frac{\sum_{i=1}^7 (x_i - \bar{x})^2}{6} \\ &\approx 7.21 \end{aligned}$$

$$\begin{aligned} \text{ii } s_n^2 &= s_7^2 \\ &= \frac{6}{7} s_6^2 \\ &\approx 6.19 \end{aligned}$$

c $E(S_6^2) = \sigma^2$ and $E(S_7^2) = \frac{6}{7}\sigma^2$.

- d** s_6^2 is the preferred estimate of σ^2 since s_6^2 is an unbiased estimate of σ^2 , whereas s_7^2 is a biased estimate of σ^2 .

EXERCISE F

- 1** Suppose random samples of size 3 of independent values are taken from a population X with mean μ and variance σ^2 .

a Show that $T_1 = \frac{4X_1 + 3X_2 + 5X_3}{12}$ is an unbiased estimator of μ .

b Show that $T_2 = \frac{2X_1 + X_2 + 3X_3}{6}$ is an unbiased estimator of μ .

- c** Which of T_1 and T_2 is the more efficient estimator? Why?

- 2 a** Consider a population with mean μ and variance σ^2 . Two independent random samples are taken with sizes 10 and 25, and the sample means $\bar{x}_{10}$ and $\bar{x}_{25}$ respectively are calculated. Which is the preferred estimate of μ ? Why?

- b** Hence explain why larger samples are better than smaller samples for estimating the population mean μ .

- 3 a** Suppose T_1 and T_2 are two independent unbiased estimators of a parameter θ . Show that $T = aT_1 + bT_2$ with $a, b \in \mathbb{R}$ is an unbiased estimator of θ if and only if $a + b = 1$.

- b** If $T_1, T_2, \dots, T_n$ are independent unbiased estimators of a parameter θ , under what condition(s)

is $T = \sum_{i=1}^n a_i T_i$, $a_i \in \mathbb{R}$ an unbiased estimator of θ ?

- 4** A population with distribution X has mean μ and variance σ^2 . Random samples $\{x_1, x_2\}$ of size 2 of independent values are taken from X .

Let $T = \lambda X_1 + (1 - \lambda) X_2$, $0 \leq \lambda \leq 1$, be an estimator of μ .

- a** Show that T is an unbiased estimator of μ .

- b** Use calculus to prove that the most efficient estimator of this form has $\lambda = \frac{1}{2}$.

- 5** Consider a population with unknown mean μ and unknown variance σ^2 . Two independent random samples $\{x_1, x_2, x_3, x_4\}$ and $\{y_1, y_2, y_3, y_4, y_5, y_6, y_7\}$ are taken.

$$\text{Let } s_X^2 = \frac{\sum_{i=1}^4 (x_i - \bar{x})^2}{3} \text{ and } s_Y^2 = \frac{\sum_{i=1}^7 (y_i - \bar{y})^2}{6}.$$

- a** Show that $t = \frac{3s_X^2 + 6s_Y^2}{9}$ is an unbiased estimate of σ^2 .

- b** If the two samples had size n and m respectively, write down an unbiased estimate t of σ^2 in

terms of $n, m, s_X^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$, and $s_Y^2 = \frac{\sum_{i=1}^m (y_i - \bar{y})^2}{m-1}$.

- 6** Consider a normal distribution X with unknown mean μ and unknown variance σ^2 . Let $\bar{X}$ be the usual sample mean estimator of μ , using random samples of n independent values.

- a** Show that $E(\bar{X}^2) > \mu^2$.
b What does the result in **a** imply?

From **Theorem 3**,

$$\text{Var}(\bar{X}) = E(\bar{X}^2) - \{E(\bar{X})\}^2$$



- 7** Suppose X and Y are independent random variables with $E(X) = \mu_X$, $\text{Var}(X) = \sigma_X^2$, $E(Y) = \mu_Y$, and $\text{Var}(Y) = \sigma_Y^2$.

A random sample of size n is taken from X , and the sample mean $\bar{x}$ and $s_X^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ are calculated.

Similarly, a sample of size m is taken from Y , and the sample mean $\bar{y}$ and $s_Y^2 = \frac{\sum_{i=1}^m (y_i - \bar{y})^2}{m-1}$ are calculated.

- a** Let $U = X + Y$.
i Find $E(U)$ and $\text{Var}(U)$. **ii** Show that $\bar{x} + \bar{y}$ is an unbiased estimate of $E(U)$.
iii Show that $s_X^2 + s_Y^2$ is an unbiased estimate of $\text{Var}(U)$.
b Let $U = aX + bY$, where $a, b \in \mathbb{R}^+$.
i Find $E(U)$ and $\text{Var}(U)$.
ii Show that $a\bar{x} + b\bar{y}$ is an unbiased estimate of $E(U)$.
iii Is $as_X^2 + bs_Y^2$ an unbiased estimate of $\text{Var}(U)$? Explain your answer.
- 8** Let $\{2, 1.5, 6.78, 4.25, 8.61, 3.2\}$ be a random sample of independent values taken from a population with unknown mean μ and unknown variance σ^2 .
a Find an unbiased estimate of μ . **b** Find an unbiased estimate of σ^2 .
c Calculate the sample variance s_n^2 .
d State the numeric relationship between your answers in **b** and **c**. Explain why we expect the value in **b** to be larger than the value in **c**.
- 9** Let $X \sim U(0, b)$ be the continuous uniform random variable with probability density function $f(x) = \begin{cases} \frac{1}{b} & 0 \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$
a Calculate explicitly $E(X)$.
b Let $\bar{X}$ be the sample mean estimator of $E(X)$, calculated from random samples of size n of independent values of X .
 Show that $2\bar{X}$ is an unbiased estimator of the parameter b , for all $n \in \mathbb{Z}^+$.

- 10** Suppose a random sample of size n of independent values is taken from a population with *known* mean μ and unknown variance σ^2 .

Show that $S_\mu^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}$ is an unbiased estimator of σ^2 .

- 11** Suppose random independent samples of independent values are taken from the same population with unknown mean μ and unknown variance σ^2 .

The variance of each sample is calculated using $s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$.

Sample A of size 4 has sample variance $s_A^2 = 3$.

Sample B of size 9 has sample variance $s_B^2 = 5$.

Sample C of size 20 has sample variance $s_C^2 = 2$.

a Let $t = \frac{4s_A^2 + 9s_B^2 + 20s_C^2}{30}$.

i Calculate the estimate t . **ii** Show that t is an unbiased estimate of σ^2 .

b If r such samples with sizes $n_1, n_2, \dots, n_r$ have sample variances $s_1^2, s_2^2, \dots, s_r^2$ respectively, suggest a formula for an unbiased estimate t of σ^2 in terms of $s_1^2, s_2^2, \dots, s_r^2$.

- 12** Let X and Y be independent random variables with means μ_X and μ_Y respectively. Let $\bar{x}$ be the sample mean of a random sample of independent values taken from X . Let $\bar{y}$ be the sample mean of a random sample of independent values taken from Y . Show that the product of sample means $\bar{x}\bar{y}$ is an unbiased estimate of $\mu_X\mu_Y$.

- 13** Consider a sample proportion $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$.

Let $q = 1 - p$ and $\hat{q} = 1 - \hat{p}$, and let n be any constant, $n \in \mathbb{Z}^+$.

a Find $E(\hat{p})$ and $E(\hat{q})$. **b** Calculate $E\left(\frac{\hat{p}\hat{q}}{n}\right)$.

c Hence explain why $\frac{\hat{p}\hat{q}}{n}$ is a biased estimate of the variance $\frac{pq}{n}$ of $\hat{p}$.

d Find an expression for an unbiased estimate of the variance $\frac{pq}{n}$ of $\hat{p}$.

G

CONFIDENCE INTERVALS FOR MEANS

It is often infeasible to calculate a population parameter. For example, consider calculating the mean weekly salary of *all* Spaniards aged 18 and over.

In the previous section we looked at ways of estimating parameters, such as the population mean μ , with approximate single values called point estimates. We want to know how confident we can be that our estimate is close to the true population parameter μ .

A **confidence interval estimate of a parameter**, in this case the population mean μ , is an interval of values between two limits together with a percentage indicating our confidence that the true parameter μ lies in that interval.

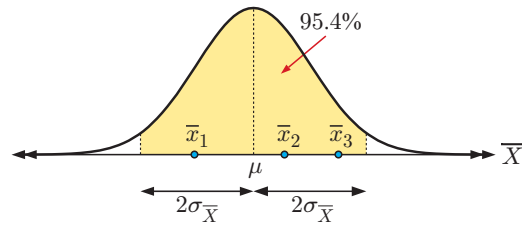
The Central Limit Theorem is used as a basis for finding the confidence intervals.

For example, consider using sample means to estimate the population mean μ .

By the CLT, we can assume that approximately 95% of the sample means from samples of size n , lie within 2 standard errors of the population mean.

$$E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

The diagram shows the distribution of sample means, $\bar{X}$.



Consider the statement “We are 95% confident that the mean weekly salary of all adult Spaniards is between 637 euros and 691 euros”.

The statement indicates that the population mean μ most likely lies in an interval between 637 euros and 691 euros. In particular, the probability that the interval contains the parameter μ is 0.95.

THE 95% CONFIDENCE INTERVAL

Consider the distribution $\bar{X}$ of all sample means $\bar{x}$ from samples of size n (large enough) taken from population X with population mean μ and population standard deviation σ . $\bar{X}$ has distribution

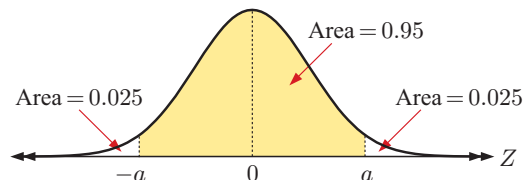
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ with mean } \mu_{\bar{X}} = \mu \text{ and standard deviation } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

The corresponding standard normal random variable is $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ and $Z \sim N(0, 1)$.

For a **95% confidence level** we need to find a for which $P(-a \leq Z \leq a) = 0.95 \dots (*)$.

Using the symmetry of the graph of the normal distribution, the statement reduces to $P(Z < -a) = 0.025$ or $P(Z < a) = 0.975$.

Using technology, we find that $a \approx 1.96$.



Therefore, in $(*)$, $P(-1.96 \leq Z \leq 1.96) = 0.95$ or $P(-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) = 0.95$.

Let $\bar{x}$ be any such sample mean.

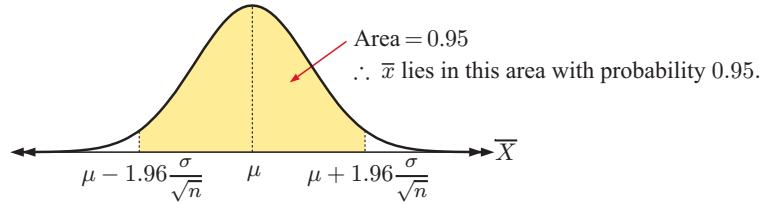
$$-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96$$

$$\therefore -1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}$$

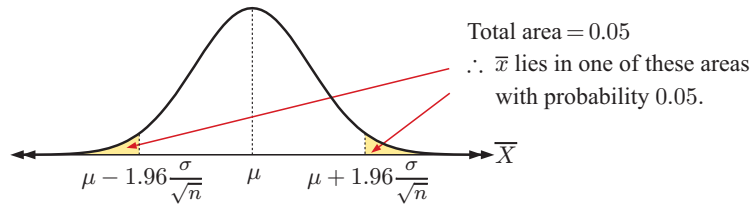
$$\therefore \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \quad \text{or equivalently} \quad \mu - 1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}$$

We therefore have the following equivalent results:

- 95% of all sample means $\bar{x}$ from samples of size n , lie between values $\mu - 1.96 \frac{\sigma}{\sqrt{n}}$ and $\mu + 1.96 \frac{\sigma}{\sqrt{n}}$ inclusive.

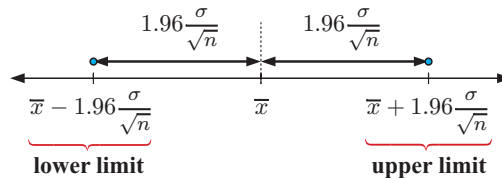


- Given the sample mean $\bar{x}$ from one sample of size n , there is probability 0.95 that $\mu - 1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}$, and probability 0.05 that $\bar{x}$ lies outside this interval.



- Given the sample mean $\bar{x}$ from one sample of size n , there is probability 0.95 that the true population mean μ satisfies $\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$, and probability 0.05 that μ lies outside this interval.

The **95% confidence interval for μ** using sample mean $\bar{x}$ is $\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$, also denoted $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ or $\left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right]$.

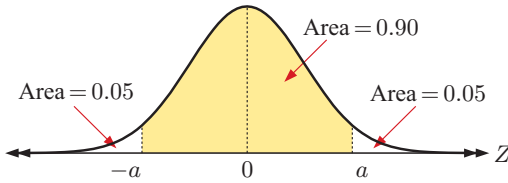


We notice that:

- The exact **centre** or midpoint of the confidence interval is the value of $\bar{x}$ for the sample taken.
- The **width** of the 95% confidence interval for μ is $2 \times 1.96 \frac{\sigma}{\sqrt{n}}$.

- The use of $\leq$ or $<$ makes no difference in the calculation of the areas and therefore the probabilities for continuous random variables, but we require $\leq$ in the definition of the confidence interval for later work on hypothesis testing.

OTHER CONFIDENCE INTERVALS FOR μ



For a 90% confidence interval, $P(Z < -a) = 0.05$ or $P(Z < a) = 0.95$.

Using technology, $a \approx 1.645$. Since a is the coefficient of $\frac{\sigma}{\sqrt{n}}$ in the confidence interval:

The 90% confidence interval for μ is $\bar{x} - 1.645 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.645 \frac{\sigma}{\sqrt{n}}$

Using this technique we obtain the following confidence intervals:

Confidence level	a	Confidence interval
90%	1.645	$\bar{x} - 1.645 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.645 \frac{\sigma}{\sqrt{n}}$
95%	1.960	$\bar{x} - 1.960 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.960 \frac{\sigma}{\sqrt{n}}$
98%	2.326	$\bar{x} - 2.326 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.326 \frac{\sigma}{\sqrt{n}}$
99%	2.576	$\bar{x} - 2.576 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.576 \frac{\sigma}{\sqrt{n}}$

The **confidence level** is the amount of confidence we place in μ being within the calculated confidence interval.



We notice that:

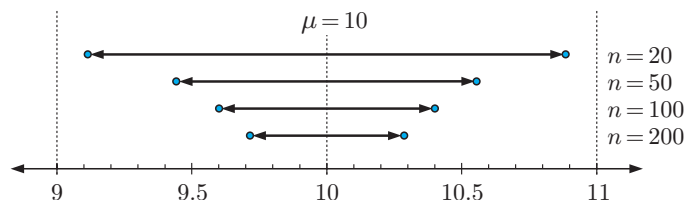
- The sample mean $\bar{x}$ is the **centre** or midpoint of the confidence interval.
- The **width** of a confidence interval is $2 \times a \times \frac{\sigma}{\sqrt{n}}$ where a is given in the table above.
- Increasing the sample size n produces confidence intervals of shorter width.

For example, consider samples of different size but all with sample mean 10 and standard deviation 2.

The 95% confidence interval is $10 - \frac{1.960 \times 2}{\sqrt{n}} \leq \mu \leq 10 + \frac{1.960 \times 2}{\sqrt{n}}$.

For various values of n we have:

n	Confidence interval
20	$9.123 \leq \mu \leq 10.877$
50	$9.446 \leq \mu \leq 10.554$
100	$9.608 \leq \mu \leq 10.392$
200	$9.723 \leq \mu \leq 10.277$



INVESTIGATION 2

CONFIDENCE LEVELS AND INTERVALS

To obtain a greater understanding of confidence intervals and levels, click on the icon. The random sampler demonstration calculates confidence intervals at various levels of your choice (90%, 95%, 98%, or 99%) and counts the intervals which include the population mean.



Example 45

A pharmaceutical company produces tablets with masses that are normally distributed with standard deviation 0.038 mg. A random sample of ten tablets was found to have mean mass 4.87 mg. Calculate a 95% confidence interval for the mean mass of these tablets, based on this sample.

Even though n is relatively small, the fact that the mass $X \sim N(\mu, (0.038)^2)$ is normally distributed ensures that $\bar{X} \sim N\left(\mu, \left(\frac{0.038}{\sqrt{10}}\right)^2\right)$ by the CLT.

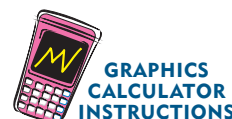
Since $\bar{x} = 4.87$, a 95% confidence interval for the mean mass μ of a tablet is

$$4.87 - 1.96 \times \frac{0.038}{\sqrt{10}} \leq \mu \leq 4.87 + 1.96 \times \frac{0.038}{\sqrt{10}}$$

which is $4.846 \leq \mu \leq 4.894$

We are 95% confident that the population mean lies in the interval $4.85 \leq \mu \leq 4.89$.

Confidence intervals can be obtained directly from your graphics calculator.



CONFIDENCE INTERVALS FOR μ WHEN σ^2 IS UNKNOWN

We usually do not know the population variance σ^2 , so instead we use s_{n-1}^2 as an unbiased estimate of σ^2 as shown in **Section F**.

If X is normally distributed then $\bar{X}$ is normally distributed, even for small sample size n .

If X is not normally distributed, for sample size n large enough, the CLT says $\bar{X}$ is approximately normally distributed.

Now if σ^2 is known, $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

If σ^2 is unknown, we use s_{n-1}^2 instead, and for $\bar{X}$ normally distributed, the random variable $T = \frac{\bar{X} - \mu}{\frac{s_{n-1}}{\sqrt{n}}}$ has a **t-distribution**, sometimes called

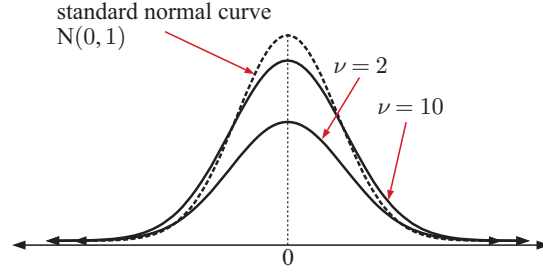
Student's t-distribution.

The Student's t -distribution is named after William Gosset who wrote under the pseudonym "Student".

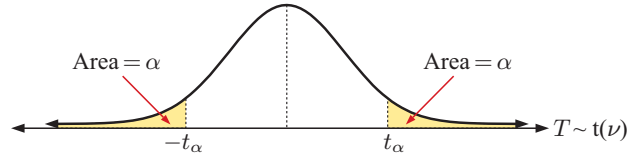


t -DISTRIBUTIONS

All t -distributions are symmetrical about the origin. They are like standardised normal bell-shaped curves, but with fatter tails. Each curve has a single parameter ν (pronounced “new”) which is a positive integer. ν is equal to the *number of degrees of freedom* of the distribution.



For a given value of ν , consider the $t(\nu)$ -distribution. Denote by t_α the value such that $P(T > t_\alpha) = \alpha$, and equivalently $P(T < -t_\alpha) = \alpha$.



Since s_{n-1}^2 has been calculated with $n-1$ degrees of freedom, we find in general that **$\nu = n - 1$** .

For example, for a sample of size 8, $\nu = 7$, and we write $T \sim t(7)$.

The graphs illustrated are those of $t(2)$, $t(10)$ and $Z \sim N(0, 1)$.

As $\nu = n - 1$ increases, the curve of the $t(n - 1)$ distribution approaches the standardised normal Z -curve.

Consider taking random samples of size n from a distribution X with mean μ and unknown variance σ^2 .

If $\bar{X}$ is normally distributed, it can be shown that $T = \frac{\bar{X} - \mu}{\frac{s_{n-1}}{\sqrt{n}}}$ follows a **t -distribution** with $n - 1$ degrees of freedom, and we write $T \sim t(n - 1)$.

In particular, when X is normally distributed with $X \sim N(\mu, \sigma^2)$, $\bar{X}$ is normally distributed for all values of n .

$\therefore T = \frac{\bar{X} - \mu}{\frac{s_{n-1}}{\sqrt{n}}}$ has the $t(n - 1)$ -distribution with $\nu = n - 1$ degrees of freedom.

Suppose n , and therefore $\nu = n - 1$, is fixed.

Since $P(-t_{0.025} \leq \frac{\bar{X} - \mu}{\frac{s_{n-1}}{\sqrt{n}}} \leq t_{0.025}) = 0.95$, the corresponding **95% confidence interval for μ** is

$$\bar{x} - \frac{s_{n-1}}{\sqrt{n}} t_{0.025} \leq \mu \leq \bar{x} + \frac{s_{n-1}}{\sqrt{n}} t_{0.025}$$

with **width** $2 \frac{s_{n-1}}{\sqrt{n}} t_{0.025}$.

Other confidence intervals can be similarly defined.

Example 46

The fat content, in grams, of 30 randomly selected pies at the local bakery was determined and recorded as:

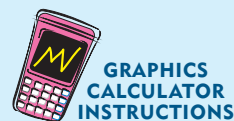
15.1 14.8 13.7 15.6 15.1 16.1 16.6 17.4 16.1 13.9 17.5 15.7 16.2 16.6 15.1
12.9 17.4 16.5 13.2 14.0 17.2 17.3 16.1 16.5 16.7 16.8 17.2 17.6 17.3 14.7

Determine a 98% confidence interval for the average fat content of all pies made.

Using technology, $\bar{x} \approx 15.897$ and $s_{n-1} \approx 1.365$.

$\bar{X}$ is approximately normally distributed by CLT, since $n = 30$ is sufficiently large.

σ is unknown and $T = \frac{\bar{X} - \mu}{\frac{s_{n-1}}{\sqrt{n}}}$ is $T \sim t(29)$.



The 98% confidence interval for μ is

$$\begin{aligned} \bar{x} - \frac{s_{n-1}}{\sqrt{n}} t_{0.01} &\leq \mu \leq \bar{x} + \frac{s_{n-1}}{\sqrt{n}} t_{0.01} \\ \therefore 15.897 - \frac{1.365}{\sqrt{30}} \times 2.462 &\leq \mu \leq 15.897 + \frac{1.365}{\sqrt{30}} \times 2.462 \\ \therefore 15.283 &\leq \mu \leq 16.511 \end{aligned}$$

Alternatively, using technology, a 98% confidence interval for μ is $15.28 \leq \mu \leq 16.51$.

Example 47

A random sample of eight independent observations of a normal random variable gave $\sum x = 72.8$ and $\sum x^2 = 837.49$. Calculate:

- an unbiased estimate of the population mean
- an unbiased estimate of the population variance, and hence an estimate of the population standard deviation
- a 90% confidence interval for the population mean.

a $\bar{x} = \frac{\sum x}{n} = \frac{72.8}{8} = 9.1$ and so 9.1 is an unbiased estimate of μ .

b $s_n^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{837.49}{8} - 9.1^2 \approx 21.876$

An unbiased estimate of σ^2 is $s_{n-1}^2 = \frac{n}{n-1} s_n^2 = \frac{8}{7} \times 21.876 \approx 25.00$

$\therefore$ an estimate of $\sigma \approx 5.00$

- c** Using $\bar{x} = 9.1$ and $s_{n-1} = 5.00$, we obtain the 90% confidence interval for μ .

$$\begin{aligned} \bar{x} - \frac{s_{n-1}}{\sqrt{n}} t_{0.05} &\leq \mu \leq \bar{x} + \frac{s_{n-1}}{\sqrt{n}} t_{0.05} \\ \therefore 9.1 - \frac{5}{\sqrt{8}} \times 1.895 &\leq \mu \leq 9.1 + \frac{5}{\sqrt{8}} \times 1.895 \\ \therefore 5.750 &\leq \mu \leq 12.45 \end{aligned}$$

Alternatively, using technology, $5.75 \leq \mu \leq 12.45$ {using the t -distribution}



DETERMINING HOW LARGE A SAMPLE SHOULD BE

When designing an experiment in which we wish to estimate the population mean, the size of the sample is an important consideration. Finding the appropriate sample size is a problem that can be solved using the confidence interval.

Example 48

Consider again **Example 46** on the fat content of pies. Suppose the population standard deviation $\sigma = 1.365$ g. How large should a sample be if we wish to be 98% confident that the sample mean will differ from the population mean by less than 0.3 grams?

We require $-0.3 < \mu - \bar{x} < 0.3$

Now the 98% confidence interval for μ is

$$\bar{x} - 2.326 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.326 \frac{\sigma}{\sqrt{n}}$$

$$\therefore -2.326 \frac{\sigma}{\sqrt{n}} \leq \mu - \bar{x} \leq 2.326 \frac{\sigma}{\sqrt{n}}$$

So, we need to find n such that

$$2.326 \frac{\sigma}{\sqrt{n}} < 0.3$$

$$\text{Consider } \sqrt{n} = \frac{2.326\sigma}{0.3} = \frac{2.326 \times 1.365}{0.3} \approx 10.583$$

$$\therefore n \approx 112.01$$

So, a sample of size at least 113 should be taken to achieve the required accuracy.

The final answer for n needs to be rounded up here.



EXERCISE G.1

- 1 The mean μ of a population is unknown, but its standard deviation is 10. In order to estimate μ , a random sample of size $n = 35$ was selected. The mean of the sample was 28.9.
 - a Find a 95% confidence interval for μ .
 - b Find a 99% confidence interval for μ .
 - c In changing the confidence level from 95% to 99%, how does the width of the confidence interval change?
- 2 When performing a statistical analysis, we can choose the confidence level for a confidence interval. Why would statisticians not always choose to use confidence intervals of at least 99%?
- 3 A random sample of size n is selected from a population with known standard deviation 11. The sample mean is 81.6.
 - a Find a 95% confidence interval for μ if: i $n = 36$ ii $n = 100$.
 - b In changing n from 36 to 100, how does the width of the confidence interval change?
- 4 The $P\%$ confidence interval for μ is $\bar{x} - a \left(\frac{\sigma}{\sqrt{n}} \right) \leq \mu \leq \bar{x} + a \left(\frac{\sigma}{\sqrt{n}} \right)$.
 If $P = 95$, then $a = 1.960$.
 Find a if P is: a 99 b 80 c 85 d 96
Hint: Use the Z -distribution.

- 5** A random sample of size $n = 50$ is selected from a population with standard deviation σ . The sample mean is 38.7.
- Find a 95% confidence interval for the mean μ if:
 - $\sigma = 6$
 - $\sigma = 15$.
 - What effect does changing σ from 6 to 15 have on the width of the confidence interval?
- 6** Neville kept records of the time that he had to wait to receive telephone support for his accounting software. During a six month period he made 167 calls and the mean waiting time was 8.7 minutes. The shortest waiting time was 2.6 minutes and the longest was 15.1 minutes.
- Estimate σ using $\sigma \approx \text{range} \div 6$. Use the normal distribution to briefly explain why this estimate for σ is a reasonable one.
 - Find a 98% confidence interval for estimating the mean waiting time for all telephone customer calls for support.
- 7** A breakfast cereal manufacturer uses a machine to deliver the cereal into plastic packets. The quality controller randomly samples 75 packets and obtains a sample mean of 513.8 grams with sample standard deviation 14.9 grams. Construct a 99% confidence interval in which the true population mean should lie.
- 8** A sample of 42 patients from a drug rehabilitation program showed a mean length of stay on the program of 38.2 days with standard deviation 4.7 days. Estimate, with a 90% confidence interval, the average length of stay for all patients on the program.
- 9** A sample of 60 yabbies was taken from a dam. The sample mean weight of the yabbies was 84.6 grams, and the sample standard deviation was 16.8 grams.
- For this yabbie population, find:
 - the 95% confidence interval for the population mean
 - the 99% confidence interval for the population mean.
 - What sample size is needed to be 95% confident that the sample mean differs from the population mean by less than 5 g?



- 10** A random sample of ten independent observations of a normal random variable gave $\sum x = 112.5$ and $\sum x^2 = 1325.31$. Calculate:
- an unbiased estimate of the population mean
 - an unbiased estimate of the population variance, and hence an estimate of the population standard deviation
 - a 90% confidence interval for the population mean.
- 11** A porridge manufacturer knows that the population variance σ^2 of the weight of contents of each packet produced is 17.8^2 grams². How many packets must be sampled to be 98% confident that the sample mean differs from the population mean by less than 3 grams?
- 12** A sample of 48 patients from an alcohol rehabilitation program showed participation time on the program had a sample variance of 22.09 days².
- Use the sample variance to estimate the population standard deviation σ .
 - How many patients would have to be sampled to be 99% confident that the sample mean number of days on the program differs from the population mean by less than 1.8 days?

PAIRED DATA (THE CASE OF MATCHED PAIRS)

Often we are interested in comparing sets of results for the same (or similarly matched) group(s) of individuals.

For example, we might consider:

- race times for a class of students at the start and finish of the athletics season
- test results for two classes of the same size, of students of similar ability.

In each case the data in the two samples obtained are matched in pairs. The two samples are not necessarily independent, for example a matched pair of race times for a particular individual at the start and finish of the athletics season. However, the individual scores in each sample must be independent for our analysis to be meaningful.

We create a **new single sample** from the **differences** of the matched pairs and proceed with our usual methods for a single sample.

For matched pairs, the population standard deviation σ will in general not be known. In such cases it is necessary to approximate σ by s_{n-1} from the sample, and use the confidence interval for the t -distribution.

Example 49

Prior to the 2004 Olympic Games an Institute of Sport took 20 elite athletes, and over a twelve month period monitored their training for the 100 m sprint. Below is the “best” time for each athlete in trials at the start and end of the year. The athletes have been recorded as the letters A to T, and times are in seconds.

Athlete	A	B	C	D	E	F	G	H	I	J
Start	10.3	10.5	10.6	10.4	10.8	11.1	9.9	10.6	10.6	10.8
End	10.2	10.3	10.8	10.1	10.8	9.7	9.9	10.6	10.4	10.6

Athlete	K	L	M	N	O	P	Q	R	S	T
Start	11.2	11.4	10.9	10.7	10.7	10.9	11.0	10.3	10.5	10.6
End	10.8	11.2	11.0	10.5	10.7	11.0	11.1	10.5	10.3	10.2

- Create i 95% ii 90% confidence intervals for the average time difference (start time – end time) for all athletes in the relevant population.
- The Institute of Sport claims their training program has improved sprint times. Do you agree? Explain your answer.

- Let $U = X_1 - X_2$, where X_1 represents the ‘start’ time and X_2 represents the ‘end’ time.

	A	B	C	D	E	F	G	H	I	J
u	0.1	0.2	-0.2	0.3	0	1.4	0	0	0.2	0.2

	K	L	M	N	O	P	Q	R	S	T
u	0.4	0.2	-0.1	0.2	0	-0.1	-0.1	-0.2	0.2	0.4

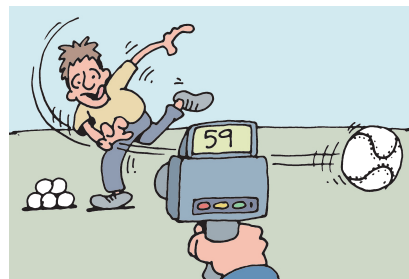
Here $n = 20$, $\bar{u} = 0.155$, and $s_{n-1}^2 = (0.344\,085\,36)^2$.

σ is unknown and $T = \frac{\bar{U} - \mu}{\frac{s_{n-1}}{\sqrt{20}}}$ is $T \sim t(19)$.

- i The 95% CI for μ is $-0.006\,04 \leq \mu \leq 0.316$.
 - ii The 90% CI for μ is $0.021\,96 \leq \mu \leq 0.2880$.
- b** An improvement in times corresponds to $\mu > 0$.
 There is sufficient evidence at the 90% level that $\mu > 0$, since $\mu = 0$ lies outside the confidence interval, and μ is given to be positive.
 There is insufficient evidence at the 95% level to suggest that $\mu > 0$, since the confidence interval contains negative values.
 Thus we would agree with the Institute of Sport's claim at the 90% level, but not at the 95% level.

EXERCISE G.2

- 1** A group of 12 year old children were asked to throw a baseball as fast as they could. A radar was used to measure the speed of each throw. One year later, the same group was asked to repeat the experiment. The results are shown below, with the children labelled A to K, and the speeds given in km h^{-1} .



Age	A	B	C	D	E	F	G	H	I	J	K
12	76	81	59	67	90	74	78	71	69	72	82
13	79	82	66	72	93	76	77	82	75	77	86

- a** Find confidence intervals for the average throwing speed difference for all children from age 12 to age 13, using a:
 - i 95% confidence level
 - ii 90% confidence level.
 - b** A sports commission report suggests that an average throwing speed difference of 5 km h^{-1} is expected between these ages. Based on these experimental results, do you agree with the sports commission report?
- 2** Pairs of identical seedlings were grown with two types of compost, one with Type 1 compost and one with Type 2. The pairs were grown side by side in various garden plots. After a period of time, the height (in cm) of each seedling was measured.

Pair	A	B	C	D	E	F	G	H
Type 1	12.1	14.6	10.1	8.7	13.2	15.1	16.5	14.6
Type 2	12.3	15.2	9.9	9.5	13.4	14.9	17.0	14.8

- a** Determine unbiased estimates of the mean and variance of the difference $d = (\text{height Type 2}) - (\text{height Type 1})$ for this paired data.
- b** Calculate a confidence interval for μ , the mean improved growth when using Type 2 compost instead of Type 1 compost, with confidence level:
 - i 95%
 - ii 99%.
- c** The manufacturer of Type 2 compost guarantees it will improve the growth of seedlings more than Type 1 compost. Comment on this claim based on your calculations.
- d** A confidence interval for μ is calculated as $[-0.032, 0.557]$.
 Find, accurate to one decimal place, the % confidence level for this confidence interval.

H

SIGNIFICANCE AND HYPOTHESIS TESTING

Visitors to the West Coast of the South Island of New Zealand are often bitten by sandflies.

According to its label, a new product claims to repel sandflies with “average protection time of more than 6 hours”. The best products currently available protect for six hours.

The government department for tourism wishes to preserve the tourist trade, and therefore needs to provide the best possible advice to tourists. How can they test the manufacturer’s claim?



HYPOTHESES

A **statistical hypothesis** is a statement about the value of a population parameter. The parameter could be a population mean μ , or a proportion p .

When a claim is made about a product, the claim can be tested statistically.

The statistician begins by formulating a **null hypothesis**, H_0 , that a parameter of the population takes a definite value, for example, that the population mean μ has value μ_0 . This statement is assumed to be true unless sufficient evidence is provided for it to be rejected. If the hypothesis is not rejected, we accept that the population mean is μ_0 , so the null hypothesis is a statement of *no difference*.

The **alternative hypothesis**, H_1 , is that there is a *difference* between μ and μ_0 . We will only accept this hypothesis if there is evidence to support it.

The statistician then gathers a random sample from the population in order to test the null hypothesis. If the test shows that H_0 should be rejected, then its alternative H_1 is accepted.

ONE-TAILED AND TWO-TAILED ALTERNATIVE HYPOTHESES

Given the null hypothesis $H_0: \mu = \mu_0$, the alternative hypothesis could be:

- $H_1: \mu > \mu_0$ (**one-tailed**)
- $H_1: \mu < \mu_0$ (**one-tailed**)
- $H_1: \mu \neq \mu_0$ (**two-tailed**, as $\mu \neq \mu_0$ could mean $\mu > \mu_0$ or $\mu < \mu_0$).

For example, consider the case of the sandfly repellent:

- If the manufacturer of the new brand wants evidence that the new product is *superior* in protection time, the hypotheses would be:

$H_0: \mu = 6$ {the new product has the same effectiveness as the old ones}

$H_1: \mu > 6$ {the new product protects for longer than the old ones}.

- If the competitor wants evidence that the new product has *inferior* protection time, the hypotheses would be:

$H_0: \mu = 6$ {the new product has the same effectiveness as the old ones}

$H_1: \mu < 6$ {the new product protects for less time than the old ones}.

The null hypothesis H_0 always states μ **equal** to a specific value.



- If a researcher studying all products on the market wants to show that the new product *differs* from the old ones, but is not concerned whether the protection time is more or less, the hypotheses would be:
 $H_0: \mu = 6$ {the new product has the same effectiveness as the old ones}
 $H_1: \mu \neq 6$ {the new product has different effectiveness from the old ones}.

ERROR TYPES

There are two types of error in decision making:

- A **Type I error** is when we make the mistake of rejecting the null hypothesis H_0 , when H_0 is in fact true.
- A **Type II error** is when we make the mistake of accepting H_0 when H_0 is in fact not true.

For example, if a coin is fair then the population proportion of *heads* it produces is $p = 0.5$.

- A Type I error would be deciding a fair coin is biased because of the event of obtaining 10 heads in 10 tosses. Although improbable, it is still possible to obtain this result with a fair coin.
- A Type II error would be accepting that a biased coin is fair (when it is in fact biased) because of the event of obtaining 7 heads in 10 tosses. This event can occur with a fair coin with reasonable probability, but the coin used may in fact be biased towards heads.

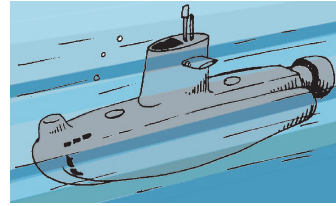
EXERCISE H.1

- 1 Explain what is meant by:
 - a Type I error
 - a Type II error
 - the null hypothesis
 - the alternative hypothesis.
- 2
 - An experimenter wishes to test $H_0: \mu = 20$ against $H_1: \mu > 20$.
 - If the mean is actually 20 but the experimenter concludes that the mean exceeds 20, what type of error has been made?
 - If the population mean is actually 21.8 but the experimenter concludes that the mean is 20, what type of error has been made?
 - A researcher wishes to test $H_0: \mu = 40$ against $H_1: \mu \neq 40$. What type of error has been made if she concludes that:
 - the mean is 40 when it is in fact 38.1
 - the mean is not 40 when it actually is 40?
- 3 In many countries where juries are used in trials, “a person is presumed innocent until proven guilty”. In this case the null hypothesis would be H_0 : the person on trial is innocent.
 - What would be the alternative hypothesis H_1 ?
 - If an innocent person is judged guilty, what type of error has been made?
 - If a guilty person is judged as innocent, what type of error has been made?
- 4 A researcher conducts experiments to determine the effectiveness of two anti-dandruff shampoos X and Y. He tests the hypotheses:

$$H_0: X \text{ and } Y \text{ have the same effectiveness} \quad H_1: X \text{ is more effective than } Y.$$
 What decision would cause:
 - a Type I error
 - a Type II error?



- 5 Current torch globes have a mean life of 80 hours. Globe Industries are considering mass production of a new globe they believe will last longer.
- If Globe Industries wants to demonstrate that their new globe lasts longer, what set of hypotheses should they consider?
 - The new globe costs less to make, so Globe Industries will adopt it unless it has an inferior lifespan to the old type. What set of hypotheses would they now consider?
- 6 The top underwater speed of submarines produced at the dockyards is 26.3 knots. The engineers modify the design to reduce drag and believe that the maximum speed will now be considerably increased. What set of hypotheses should they consider to test whether or not the new design has produced a faster submarine?



HYPOTHESIS TESTING FOR THE MEAN WITH ONE SAMPLE WHERE σ IS KNOWN

Consider a population X with unknown mean μ and known standard deviation σ .

Suppose we take a random sample of size n (large) of independent values of X , and calculate the sample mean $\bar{x}$.

We wish to test whether or not X has mean μ equal to a specific value μ_0 , based on this one sample mean $\bar{x}$. The test can be performed as a two-tailed test, or as a one-tailed test.

In either case we begin by assuming the **null hypothesis** $H_0: \mu = \mu_0$. If this holds, then $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$, and therefore $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ has distribution $N(0, 1)$, called the **Null distribution**.

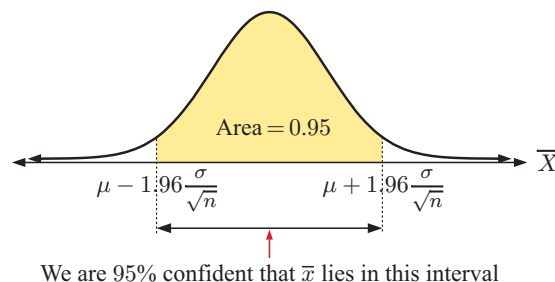
For the given sample, we calculate $z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$, called the **test statistic**.

Given the distributions for $\bar{X}$ and Z , we can calculate probabilities for where we expect a sample mean $\bar{x}$ to lie.

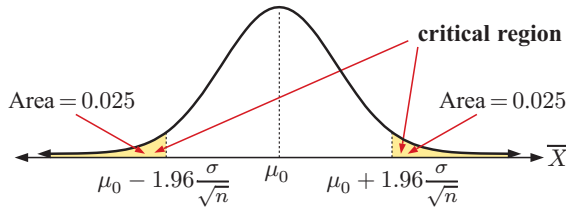
If $\bar{x}$ lies in an extreme outer tail of the $\bar{X}$ -distribution, or equivalently z^* lies in an extreme outer tail of the Z -distribution, then either we have an extremely unlikely sample, or else the null hypothesis is incorrect. In this case we need to make a decision on whether or not to reject H_0 and accept the alternative hypothesis H_1 .

EXAMPLE TWO-TAILED TEST

Consider a two-tailed test with null hypothesis $H_0: \mu = \mu_0$ and alternative hypothesis $H_1: \mu \neq \mu_0$. Assuming the population mean is indeed μ_0 , we know from the 95% confidence interval that for 95% of samples of size n , the sample mean $\bar{x}$ will satisfy $\mu_0 - 1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + 1.96 \frac{\sigma}{\sqrt{n}}$.



For the remaining 5% of samples, $\bar{x}$ will lie outside the interval $[\mu_0 - 1.96 \frac{\sigma}{\sqrt{n}}, \mu_0 + 1.96 \frac{\sigma}{\sqrt{n}}]$ in the outer tails of the distribution. We call these tails the **critical region** of the $\bar{X}$ -distribution, with boundary **critical values** $\mu_0 - 1.96 \frac{\sigma}{\sqrt{n}}$ and $\mu_0 + 1.96 \frac{\sigma}{\sqrt{n}}$ as shown:



2.5% of sample means lie in the left section of the critical region.
2.5% of sample means lie in the right section of the critical region.



If our particular sample mean $\bar{x}$ lies in the critical region, we can be 95% confident that μ_0 is not the population mean. There is 5% uncertainty, so we make the decision, at the 5% **level of significance**, to reject the null hypothesis H_0 and accept the alternative hypothesis H_1 .

We note that there is probability 0.05 of making a Type I error. Although improbable, it is still possible that μ_0 is the true population mean, and we happened to select a random sample whose mean $\bar{x}$ had very low probability.

EXAMPLE ONE-TAILED TESTS

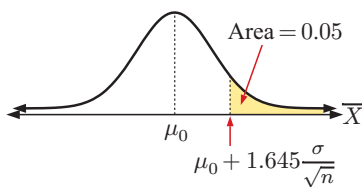
In the case of a one-sided alternative hypothesis, the critical region will be one tail only, and there is only one critical value.

To calculate the critical value for a 5% level of significance, we need to find k such that $P(Z \geq k) = 0.05$. We find $k \approx 1.645$.

Hence, for a 5% level of significance, we have:

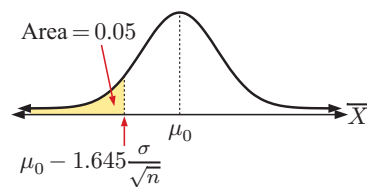
One-tailed (right) test

$$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0$$



One-tailed (left) test

$$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0$$



The level of significance determines the area of the critical region and therefore the critical values.

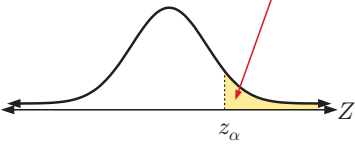
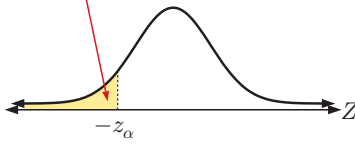
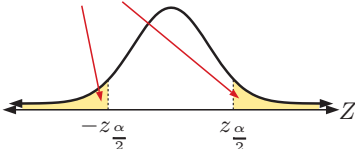


TESTING PROCEDURE

We have seen that the distribution of the test statistic $z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ is $Z \sim N(0, 1)$, called the **null distribution**.

CRITICAL REGIONS

For a level of significance α which is the area of the critical region, we can calculate the critical value(s):

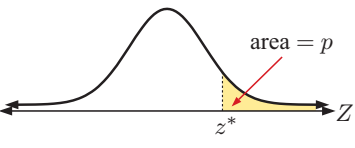
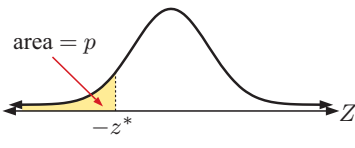
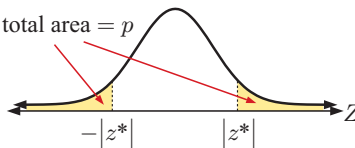
One-tailed (right) test	One-tailed (left) test	Two-tailed test
$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0$	$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0$	$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0$
$\alpha = \text{area of critical region}$	$\alpha = \text{area of critical region}$	$\alpha = \text{total area of critical region}$
		
Find z_α by solving $P(Z \geq z_\alpha) = \alpha$.	Find $-z_\alpha$ by solving $P(Z \leq -z_\alpha) = \alpha$.	Find $z_{\frac{\alpha}{2}}$ by solving $P(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.
For example, $z_{0.01} \approx 2.326$	For example, $-z_{0.01} \approx -2.326$	For example, $z_{\frac{\alpha}{2}} \approx 2.576$ and $-z_{\frac{\alpha}{2}} \approx -2.576$

In each case,

$\bar{x}$ lies in the critical region of the $\bar{X}$ -distribution $\Leftrightarrow z^*$ lies in the critical region of the null distribution.

p-VALUES

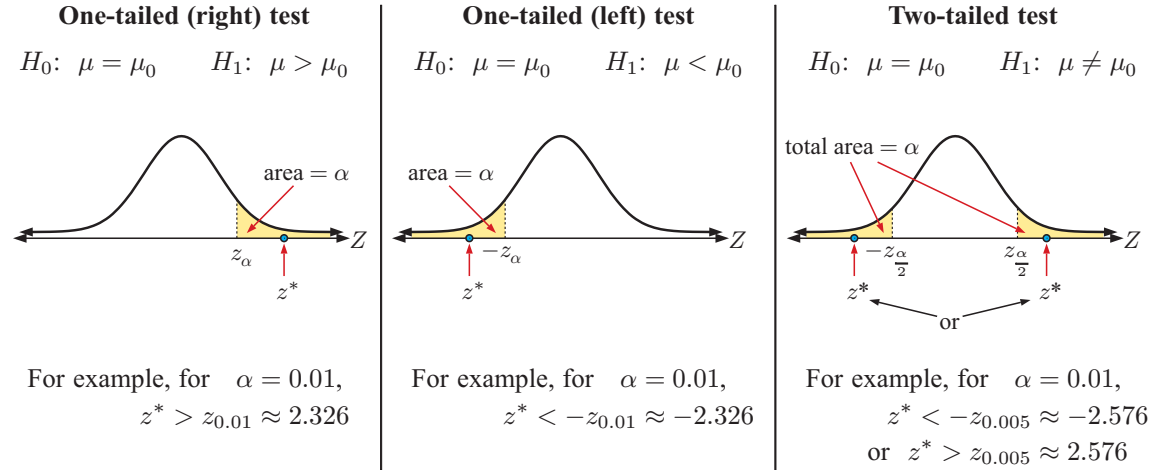
Given our one sample mean $\bar{x}$ and corresponding test statistic z^* , we define the **p-value** to be the following probability:

One-tailed (right) test	One-tailed (left) test	Two-tailed test
$p = P(Z \geq z^*)$	$p = P(Z \leq z^*)$	$p = P(Z \geq z^*) + P(Z \leq - z^*)$ $= 2P(Z \geq z^*)$ {by symmetry}
		

DECISION MAKING

We decide to **reject H_0 in favour of H_1** if any of the following (equivalent) properties hold:

- (1)** The test-statistic $z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ lies in the critical region of the null distribution $Z \sim N(0, 1)$

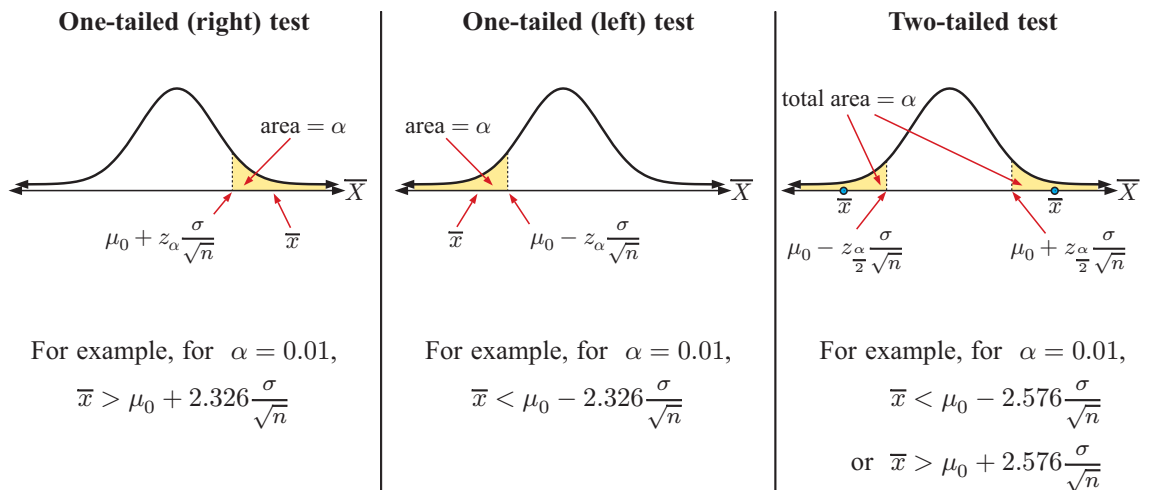


- (2)** The p -value is strictly less than α

This is a comparison of the area of the tail(s) defined by z^* (or equivalently the area of the tail(s) defined by $\bar{x}$), with the area of the critical region.

One-tailed (right) test	One-tailed (left) test	Two-tailed test
Reject if $p = P(Z \geq z^*) < \alpha$	Reject if $p = P(Z \leq z^*) < \alpha$	Reject if $p = 2P(Z \geq z^*) < \alpha$

- (3)** $\bar{x}$ lies in the critical region of the $\bar{X}$ -distribution



Otherwise, we do not reject H_0 as there is insufficient evidence to reject it. Note that although we accept H_0 , we have not actually *proved* H_0 . Rather, we simply have not found sufficient evidence against H_0 .

HYPOTHESIS TESTS AND CONFIDENCE INTERVALS

For **two-tailed** tests, we can check the result of the hypothesis test using the appropriate confidence interval.

For example,

For a two-tailed hypothesis test at the 5% level of significance, we accept H_0 if and only if μ_0 lies within the 95% confidence interval for μ .

This rule is *not* true for one-tailed tests.



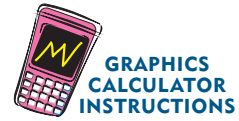
USING A GRAPHICS CALCULATOR

Click on the icon to obtain instructions for **TI** and **Casio** calculators.

Be aware that your calculator may use different notation to that used in IB.

For example:

- with **Casio** calculators, s_{n-1} is s_x
- with **TI** calculators, s_{n-1} is S_x .



REPORTING ON A HYPOTHESIS TEST

There are effectively 7 steps in reporting on a hypothesis test:

- | | |
|-------------------------------|--|
| (1) Hypotheses: | State the null and alternative hypotheses. (Specify whether it is a one- or two-tailed test.) |
| (2) Null distribution: | State the null distribution of the test statistic. |
| (3) Test statistic: | Calculate the test statistic from the sample evidence. |
| (4) Decision rule: | State the decision rule based on the significance level α . |
| (5) Evidence: | Find the p -value using your graphics calculator <i>or</i> find the critical values and the critical region. |
| (6) Decision: | Make your decision to reject or not reject H_0 , based on the significance level. |
| (7) Conclusion: | Write a brief conclusion giving your decision some contextual meaning. |

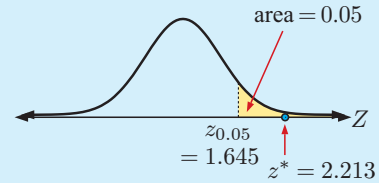
Example 50

The manager of a restaurant chain goes to a seafood wholesaler and inspects a large catch of over 50 000 prawns. It is known that the population standard deviation is 4.2 grams. She will buy the catch if the mean weight exceeds 55 grams per prawn. A random sample of 60 prawns is taken, and the mean weight is 56.2 grams. Is there sufficient evidence at a 5% level to reject the catch?

Suppose the weight of prawns has distribution X with mean μ and $\sigma = 4.2$ g.

$$\therefore \bar{X} \sim N\left(\mu, \frac{(4.2)^2}{60}\right)$$

(1) Hypotheses:	$H_0: \mu = 55$	$H_1: \mu > 55$ (one-tailed test)
(2) Null distribution:	Z -distribution ($\sigma = 4.2$ g is known)	
(3) Test statistic:	$z^* = \frac{56.2 - 55}{\frac{4.2}{\sqrt{60}}} \approx 2.213$	
(4) Decision rule:	Reject H_0 if the p -value is less than 0.05.	or Reject H_0 if z^* lies in the critical region of Z .
(5) Evidence:	$p\text{-value} = P(Z \geq 2.213) \approx 0.0134$	or
(6) Decision:	Since the p -value is less than 0.05, we reject H_0 .	or Since z^* lies in the critical region, we reject H_0 .
(7) Conclusion:	Sufficient evidence exists at the 5% level to accept H_1 , that the mean weight exceeds 55 grams. So, on this evidence, the manager should purchase the catch.	



EXERCISE H.2

- Show that if $\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$
then $\mu - 1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}$.
- Find z_α and $z_{\frac{\alpha}{2}}$ for: **a** $\alpha = 0.05$ **b** $\alpha = 0.01$
- A population has known variance $\sigma^2 = 15.79$. A sample of size 36 is taken and the sample mean $\bar{x} = 23.75$. We are required to test the hypothesis $H_0: \mu = 25$ against $H_1: \mu < 25$.
 - Find:
 - the test statistic
 - the null distribution
 - the p -value.
 - What decision should be made at a 5% level using:
 - the test statistic
 - the p -value?
- For each of the following hypotheses, find the critical region for the test statistic for the standard normal distribution with $n \geq 30$ and level of significance: **i** $\alpha = 0.05$ **ii** $\alpha = 0.01$
 - $H_0: \mu = 40$
 $H_1: \mu > 40$
 - $H_0: \mu = 50$
 $H_1: \mu < 50$
 - $H_0: \mu = 60$
 $H_1: \mu \neq 60$
- A statistician believes that a population which has a standard deviation of 12.9, has a mean μ that is greater than 80. To test this, he takes a random sample of 200 measurements, and the sample mean is 83.1. He then performs a hypothesis test with significance level $\alpha = 0.01$.
 - Write down the null and alternative hypotheses.
 - State the null distribution.
 - State the decision rule.
 - Make a decision to reject or not reject H_0 .
 - Find the value of the test statistic.
 - Find and illustrate the critical region.
 - State the conclusion for the test.

- 6 Bags of salted cashew nuts display net contents 100 g. The manufacturer knows that the standard deviation of the population is 1.6 g.

A customer claims that the bags have been lighter in recent purchases, so the factory quality control manager decides to investigate. He samples 40 bags and finds that their mean weight is 99.4 g.

Perform a hypothesis test at the 5% level of significance, using critical regions, to determine whether the customer's claim is valid.

- 7 An alpaca breeder wants to produce fleece which is extremely fine. In 2008, his herd had mean fineness 22.3 microns with standard deviation 2.89 microns. The standard deviation remains relatively constant over time. In 2012, a sample of 80 alpacas from the herd was randomly selected, and the mean fineness was 21.2 microns.

- Perform a two-tailed hypothesis test at the 5% level of significance, using p -values to determine whether the herd fineness has changed.
- Use a 95% confidence interval to check the result of your test.



A machine packs sugar into 1 kg bags. It is known that the masses of the bags of sugar are normally distributed with a variance 2.25 g. A random sample of eight filled bags was taken and the masses of the bags measured to the nearest gram. Their masses in grams were: 1001, 998, 999, 1002, 1001, 1003, 1002, 1002. Perform a test at the 1% level, to determine whether the machine overfills the bags.

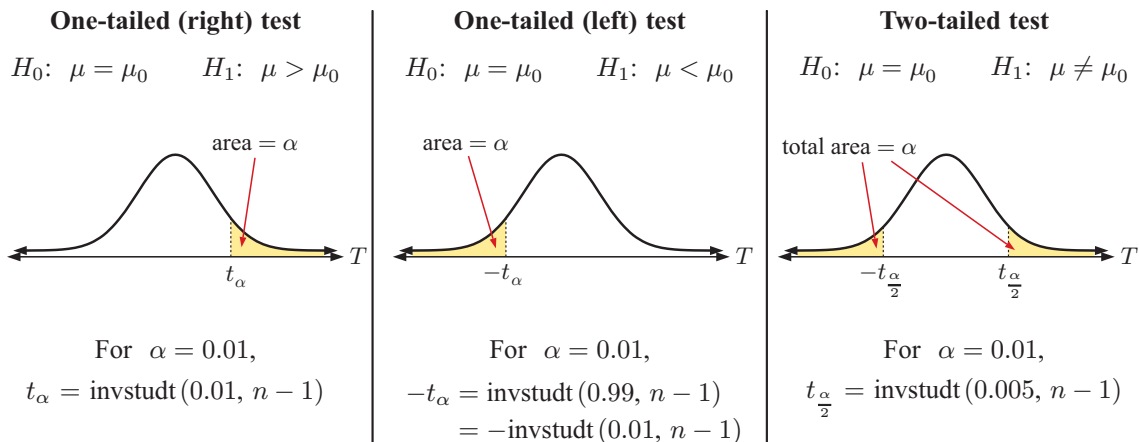
HYPOTHESIS TESTS WHERE THE POPULATION VARIANCE σ^2 IS UNKNOWN

With σ^2 unknown, we can still define the null and alternative hypotheses as before. The distribution of $\bar{X}$ is $N\left(\mu_0, \frac{\sigma^2}{n}\right)$.

However, since σ^2 is unknown, we use the unbiased estimate s_{n-1}^2 of σ^2 .

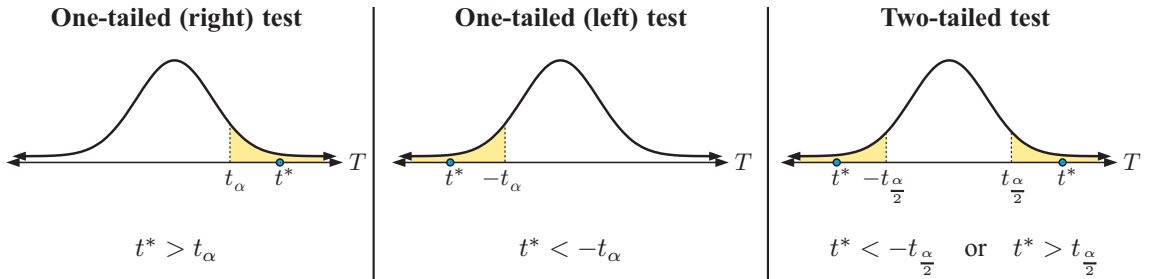
We define the **test statistic** $t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}}$ which lies in the distribution $t(n-1)$ called the **null distribution** for this case.

The **critical region** for t^* and associated **critical values** in the $t(n-1)$ -distribution need to be calculated for the given value of n and given significance level α .



We decide to **reject H_0 in favour of H_1** if either of the following (equivalent) properties hold:

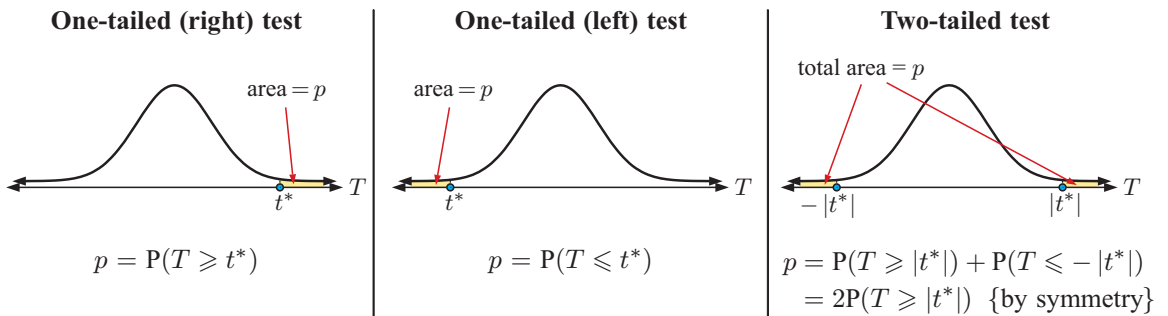
- (1)** The test statistic $t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}}$ lies in the critical region of the $t(n-1)$ -distribution for the given level of significance α :



- (2)** The p -value is strictly less than α

This is a comparison of the area of the tail(s) defined by t^* with the area of the critical region.

The p -value in this case is defined as follows:



Otherwise, we do not reject H_0 as there is insufficient evidence to do so, and so we accept H_0 .

Example 51

In 2010, the average house price in a suburb was \$235 000. In 2012, a random sample of 200 houses in the suburb was taken. The sample mean was $\bar{x} = \$215\,000$, and an estimate of the standard deviation was $s_{n-1} = \$30\,000$.

Is there evidence at the 5% level that the average house price has changed?

Suppose the price of a house has distribution X with mean μ and σ^2 unknown.

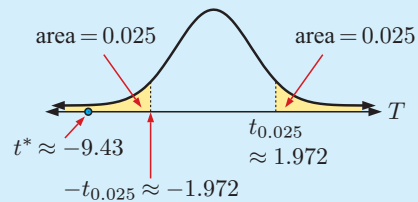
$$\therefore \bar{X} \sim N\left(\mu, \frac{\sigma^2}{200}\right).$$

- (1) Hypotheses:** $H_0: \mu = 235\,000$ $H_1: \mu \neq 235\,000$ (two-tailed test)
- (2) Null distribution:** t -distribution with $\nu = 199$, $s_{n-1} = 30\,000$ (σ^2 is unknown)
- (3) Test statistic:** $t^* = \frac{\bar{x} - \mu}{\frac{s_{n-1}}{\sqrt{n}}} = \frac{215\,000 - 235\,000}{\frac{30\,000}{\sqrt{200}}} \approx -9.428$ with 199 degrees of freedom.
- (4) Decision rule:** Reject H_0 if the p -value is less than 0.05. or Reject H_0 if t^* lies in the critical region of t .

(5) Evidence:

$$\begin{aligned}
 & p\text{-value} \\
 &= P(T \geq 9.43) \\
 &\quad + P(T \leq -9.43) \\
 &\approx 1.11 \times 10^{-17}
 \end{aligned}$$

or

**(6) Decision:**

Since the p -value is less than 0.05, we reject H_0 .

or

Since t^* lies in the critical region, we reject H_0 .

(7) Conclusion:

Sufficient evidence exists at the 5% level of significance, to suggest that $\mu \neq \$235\,000$. We conclude that the average house price in 2012 was different to the average house price in 2010.

Example 52

Fabtreed manufacture motorcycle tyres. Under normal test conditions, the average stopping time for motorcycles travelling at 60 km/h is 3.12 seconds. The production team have recently designed and manufactured a new tyre tread. They took 41 stopping time measurements under the usual test conditions, and found that the mean time was 3.03 seconds with sample standard deviation 0.27 seconds.

Is there sufficient evidence, at a 1% level, to support the team's belief that they have improved the stopping time?

Suppose the motorcycle stopping time with the new tread has distribution X with mean μ and unknown σ .

$$\therefore \bar{X} \sim N\left(\mu, \frac{\sigma^2}{41}\right).$$

(1) Hypotheses:

$$H_0: \mu = 3.12 \quad H_1: \mu < 3.12 \quad (\text{one-tailed test})$$

(2) Null distribution:

$$t\text{-distribution with } \nu = 40, s_n^2 = 0.27^2 \quad (\sigma^2 \text{ is unknown})$$

(3) Test statistic:

$$s_{n-1}^2 = \frac{n}{n-1} \times s_n^2 = \frac{41}{40} \times 0.27^2 \approx 0.074\,72$$

$$\therefore s_{n-1} \approx 0.273\,35$$

$$\therefore t^* = \frac{3.03 - 3.12}{\frac{0.273\,35}{\sqrt{41}}} \approx -2.108$$

(4) Decision rule:

Reject H_0 if the p -value is less than 0.01.

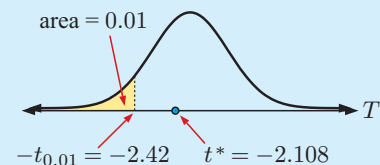
or

Reject H_0 if t^* lies in the critical region of t .

(5) Evidence:

$$\begin{aligned}
 p\text{-value} &= P(T \leq -2.108) \\
 &\approx 0.020\,67
 \end{aligned}$$

or

**(6) Decision:**

Since the p -value is not less than 0.01, we do not reject H_0 .

or

Since t^* is not in the critical region, we do not reject H_0 .

(7) Conclusion: There is insufficient evidence for us to reject H_0 , so at the 1% level of significance, we retain H_0 . We conclude that there is not an improvement in stopping time due to the new tread pattern.

EXERCISE H.3

- 1 Suppose $\nu = 15$. Find t_α and $t_{\frac{\alpha}{2}}$ for:
 - a $\alpha = 0.05$
 - b $\alpha = 0.01$
- 2 A population has unknown variance σ^2 . A sample of size 24 is taken and the sample mean $\bar{x} = 17.14$ with standard deviation 4.365. We are required to test the hypothesis $H_0: \mu = 18.5$ against $H_1: \mu \neq 18.5$.
 - a Find:
 - i the test statistic
 - ii the null distribution
 - iii the p -value.
 - b What decision should be made at a 5% level using:
 - i the test statistic
 - ii the p -value?
- 3 A liquor store claimed that the mean price of a bottle of wine had fallen from what it was 12 months previously. Records show that 12 months ago the mean price was \$13.45 for a 750 mL bottle. A random sample of prices of 389 different bottles of wine is now taken from the store. The mean price for the sample is \$13.30, with sample standard deviation \$0.25. Is there sufficient evidence at a 2% level to reject the claim? In your answer state:
 - a the null and alternative hypotheses
 - b the null distribution
 - c the test statistic
 - d the p -value
 - e your conclusion.
- 4 A machine is used to fill bottles with 500 mL of water. Ten random measurements of the volumes put in different bottles give a mean of 499 mL with standard deviation 1.2 mL. Assuming that the volumes of water are normally distributed, test at the 1% level whether there is a significant difference from the expected value.
- 5 While peaches are being canned, 250 mg of preservative is supposed to be added by a dispensing device. To check the machine, the quality controller obtains 60 random samples of dispensed preservative. She finds that the mean preservative added was 242.6 mg with sample standard deviation 7.3 mg.
 - a At a 5% level, is there sufficient evidence that the machine is not dispensing with mean 250 mg? Set out your solution in full, giving either a p -value or a critical value, and state your decision.
 - b Use a confidence interval to verify your answer.
- 6 Free range chickens are found to have mean meat protein content of 24.9 units per kg. 50 chickens are randomly chosen from a battery cage. These chickens are fed special meals which are supposed to increase their protein content. Following the feeding program, the sample had mean meat protein content 26.1 units/kg with standard deviation 6.38 units/kg.
At a 5% level of significance, test the claim that the chickens on the feeding program have a higher meat protein content.



- 7 The management of a golf club claimed that the mean income of its members was in excess of €95 000, so its members could afford to pay increased annual subscriptions. To show that this claim was invalid, the members sought the help of a statistician. The statistician was to examine the current tax records of a random sample of members fairly, and test the claim at a 0.02 significance level. The statistician found, from his random sample of 113 club members, that the average income was €96 318 with standard deviation €14 268.



- a Find an unbiased estimate of the population standard deviation.
- b State the null and alternative hypotheses when testing this claim.
- c State the null distribution.
- d Find the test statistic.
- e Find the p -value when testing the null hypothesis.
- f Find the critical region for rejection of the null hypothesis, and sketch it.
- g State whether or not there is sufficient evidence to reject management's claim.
- h Would the statistician be committing a Type I or Type II error if his assertion was incorrect?
- i Find a 99% confidence interval for the mean income of members, and comment on your result. Explain why we check with a 99% confidence interval.

MATCHED PAIRS

Example 53

Consider again the Institute of Sport problem in **Example 49** on page 98.

Prior to the 2004 Olympic Games an Institute of Sport took 20 elite athletes, and over a twelve month period monitored their training for the 100 m sprint. Below is the “best” time for each athlete in trials at the start and end of the year. The athletes have been recorded as the letters A to T, and times are in seconds.

Athlete	A	B	C	D	E	F	G	H	I	J
Start	10.3	10.5	10.6	10.4	10.8	11.1	9.9	10.6	10.6	10.8
End	10.2	10.3	10.8	10.1	10.8	9.7	9.9	10.6	10.4	10.6

Athlete	K	L	M	N	O	P	Q	R	S	T
Start	11.2	11.4	10.9	10.7	10.7	10.9	11.0	10.3	10.5	10.6
End	10.8	11.2	11.0	10.5	10.7	11.0	11.1	10.5	10.3	10.2

Conduct a hypothesis test at the 5% level of significance, to determine whether the program has significantly improved the athletes' performance.

Let $U = X_1 - X_2$ where X_1 represents the time before and X_2 represents the time after the program.

	A	B	C	D	E	F	G	H	I	J
u	0.1	0.2	-0.2	0.3	0	1.4	0	0	0.2	0.2

	K	L	M	N	O	P	Q	R	S	T
u	0.4	0.2	-0.1	0.2	0	-0.1	-0.1	-0.2	0.2	0.4

Here $n = 20$, $\bar{u} = 0.155$, and $s_{n-1} = 0.344\,085\,36$.

(1) Null hypotheses: $H_0: \mu = 0$ (times have not improved)

$H_1: \mu > 0$ (one-tailed test as testing to see if times have improved)

(2) Null distribution: t -distribution (σ^2 is unknown)

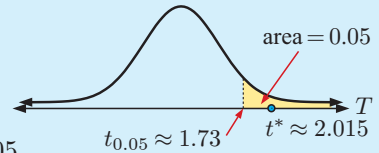
(3) Test statistic: $t^* = \frac{\bar{u} - \mu}{\frac{s_{n-1}}{\sqrt{n}}} = \frac{0.155 - 0}{0.0769} \approx 2.014\,56$

(4) Decision rule: Reject H_0 if p -value is less than 0.05

(5) Evidence: $p\text{-value} \approx P(T \geq 2.014\,56) \approx 0.029\,16$

(6) Decision: Since the p -value is less than 0.05, we reject H_0 . We observe that the test statistic $t^* \approx 2.014\,56$ lies inside the critical region.

(7) Conclusion: There is sufficient evidence at the 5% level, to conclude that the sprint times of the athletes have improved after the implementation of the program.



We have rejected the null hypothesis, yet the 95% confidence interval for μ does contain the value $\mu = 0$. This is because we have a one-tailed test.



EXERCISE H.4

- 1 A mathematics coaching school claims to significantly increase students' test results over a period of several coaching sessions. To test their claim a teacher tested 12 students prior to receiving coaching and recorded their results. The students were not given the answers or their results. At the conclusion of the coaching, the teacher gave the same test as before to check on the improvement. The paired results were:

Student	A	B	C	D	E	F	G	H	I	J	K	L
Before coaching	15	17	25	11	28	20	23	34	27	14	26	26
After coaching	20	16	25	18	28	19	26	37	31	13	27	20

Conduct a hypothesis test at a 5% level of significance, to see if the school's claim was true.

- 2 Consider again the baseball problem on page 99:

A group of 12 year old children were asked to throw a baseball as fast as they could. A radar was used to measure the speed of each throw. One year later, the same group was asked to repeat the experiment. The results are shown below, with the children labelled A to K, and the speeds given in km h^{-1} .

Age	A	B	C	D	E	F	G	H	I	J	K
12	76	81	59	67	90	74	78	71	69	72	82
13	79	82	66	72	93	76	77	82	75	77	86

A sports commission report suggests that an average throwing speed difference of 5 km h^{-1} is expected between these ages. Conduct a hypothesis test at a 5% level of significance to determine if the report's claim is valid.

3 Consider again the seedling problem on page 99:

Pairs of identical seedlings were grown with two types of compost, one with Type 1 compost and one with Type 2. The pairs were grown side by side in various garden plots. After a period of time, the height (in cm) of each seedling was measured.

Pair	A	B	C	D	E	F	G	H
Type 1	12.1	14.6	10.1	8.7	13.2	15.1	16.5	14.6
Type 2	12.3	15.2	9.9	9.5	13.4	14.9	17.0	14.8

The manufacturer of Type 2 compost guarantees it will improve the growth of seedlings more than Type 1 compost. Based on these experimental results, conduct a hypothesis test at a 5% level, to determine whether the claim is valid.

THE PROBABILITY OF ERROR

In our hypothesis tests, we have probability α that for $\mu = \mu_0$, our sample mean $\bar{x}$ could (validly) lie in the critical region of the $\bar{X}$ -distribution.

The **probability of making a Type I error** = P(rejecting H_0 when H_0 is in fact true)
= α , the level of significance.

Given that P(Type I error) = α , we let P(Type II error) = β .

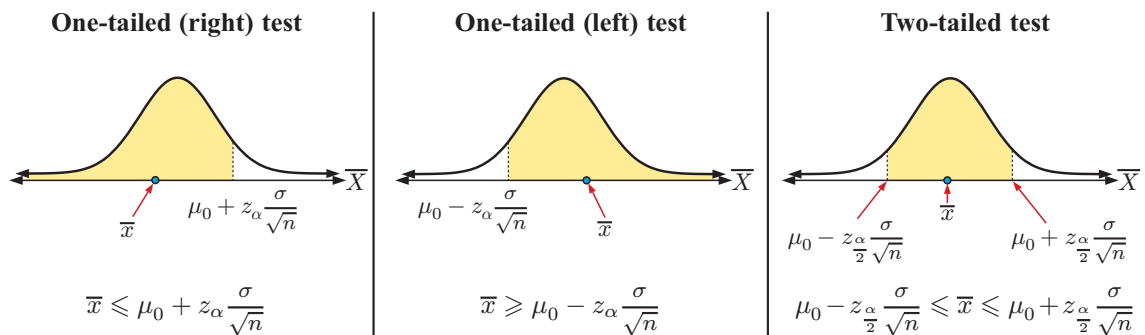
In **Example 52**, the null hypothesis H_0 is accepted since there is insufficient evidence to reject H_0 . There is the chance here of making a Type II error, which is accepting H_0 when H_0 is in fact not true. If the significance level was $\alpha = 0.05$ and not 0.01, then H_0 would have been rejected. Thus β is dependent on α .

By reducing α = P(Type I error), the probability β = P(Type II error) is increased.

In fact: We can only calculate β = P(Type II error) if μ , σ , and n are all known.

Suppose we conduct a hypothesis test with level of significance α and null hypothesis $H_0: \mu = \mu_0$, and suppose σ^2 is known.

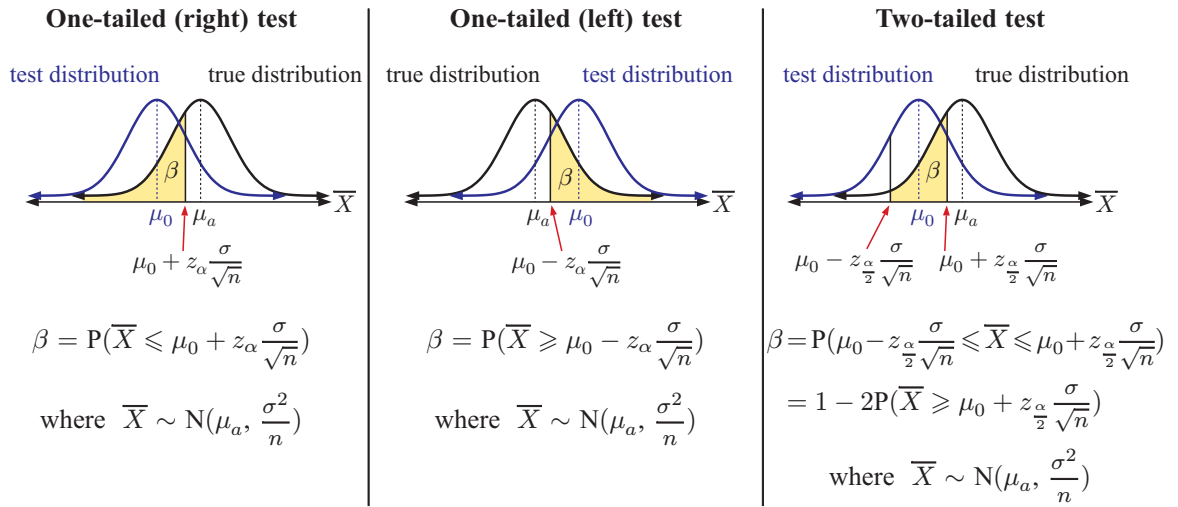
We accept H_0 if the sample mean $\bar{x}$ does not lie in the critical region of the $\bar{X}$ distribution:



But suppose H_0 is false, and the true population mean is actually $\mu = \mu_a$, where $\mu_a \neq \mu_0$.

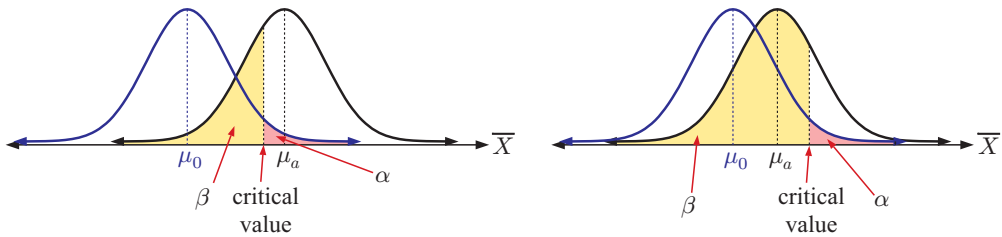
In this case, β = P(Type II error)

= P(accepting $H_0: \mu = \mu_0 \mid \mu = \mu_a$) is calculated as follows:



The **power** of a test is defined to be $1 - \beta$, which equals the probability of (correctly) rejecting the null hypothesis H_0 when H_0 is false.

Note that the closer μ_a is to μ_0 , that is the smaller $|\mu_a - \mu_0|$ is, the larger the value of $\beta = P(\text{Type II error}) = P(\text{accept } H_0 \mid H_0 \text{ is false})$, and hence the lower the power of the test.



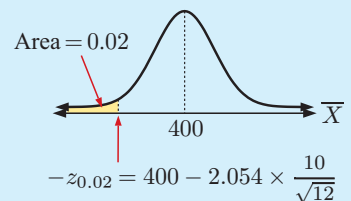
Example 54

Cans of chickpeas are labelled 400 g and it is known that the true weight of cans is normally distributed with a standard deviation of 10 g.

A statistician wishes to conduct a test to see if the mean weight of a can is less than 400 g. He uses a sample of 12 cans, and a 2% level of significance.

- Find the probability of:
 - a Type I error
 - a Type II error, given the true mean $\mu = 395$ g.
- Find the power of the test, given the true mean $\mu = 395$ g.

- $P(\text{Type I error}) = \alpha = 0.02$
 - $H_0: \mu = 400 \text{ g} \quad H_1: \mu < 400 \text{ g}$
 H_0 is retained if $\bar{x} \geq 400 - 2.054 \times \frac{10}{\sqrt{12}} \approx 394.0713$
 $\therefore \beta = P(\text{Type II error})$
 $= P(\bar{X} \geq 394.0713)$
 ≈ 0.626
- Power $= 1 - \beta \approx 0.374$



Example 55

A sample of size 25 is taken from a normal population with unknown mean μ and known variance 36.

To test the hypotheses: $H_0: \mu = 42$, $H_1: \mu > 42$ the decision rule is:

accept H_0 if the sample mean $\bar{x} \leq 43.5$

reject H_0 if the sample mean $\bar{x} > 43.5$.

- a Find the probability of a Type I error for the given decision rule with critical value 43.5.
- b Suppose the true value of the mean is 44.9.
 - i Find the probability of a Type II error.
 - ii Suppose the critical value in the decision rule is changed. For what critical value is $P(\text{Type I error}) = P(\text{Type II error})$?

a $\bar{X} \sim N(\mu, \frac{36}{25})$

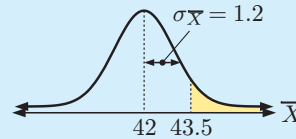
$\therefore$ under the null hypothesis $\bar{X} \sim N(42, 1.2^2)$

$\therefore P(\text{Type I error})$

$= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})$

$= P(\bar{X} > 43.5)$

≈ 0.106



b i If $\mu = 44.9$, then $\bar{X} \sim N(44.9, 1.2^2)$.

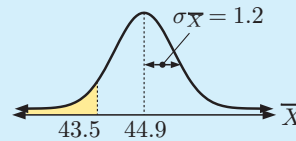
We accept H_0 if $\bar{x} \leq 43.5$.

$\therefore P(\text{Type II error})$

$= P(\text{accepting } H_0 \text{ when } H_0 \text{ is false})$

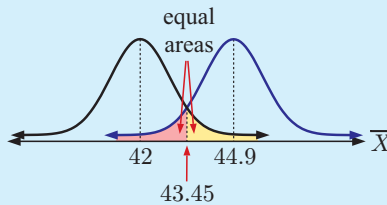
$= P(\bar{X} \leq 43.5)$

≈ 0.122



- ii By symmetry, since the standard deviation for $\bar{X}$ is constant at 1.2 for both the test distribution and the true distribution, $P(\text{Type I error}) = P(\text{Type II error})$ for the critical value which is the average of 42 and 44.9.

$$\text{The critical value} = \frac{42 + 44.9}{2} = 43.45$$



$$P(\text{Type I error}) = P(\text{Type II error}) \approx 0.113$$

EXERCISE H.5

- 1** In a population of adult trout, it is known that the length X cm is normally distributed with known variance 6 cm^2 , but the mean μ cm is unknown.
It is proposed to test the hypotheses $H_0: \mu = 27 \text{ cm}$
 $H_1: \mu > 27 \text{ cm}$
using the mean $\bar{x}$ of a sample of size 9.
- a** Find the decision rule for the test, in terms of $\bar{x}$, that corresponds to a significance level of
 - i** 5% **ii** 1%.
 - b** Suppose the true value of the mean is $\mu = 29.2$.
Calculate $P(\text{Type II error})$ when $P(\text{Type I error})$ is
 - i** 0.05 **ii** 0.01.
- 2** A sample of size 16 is taken from a normal population with unknown mean μ and known variance 64. The sample is used to test the hypotheses $H_0: \mu = 150$, $H_1: \mu > 150$.
The decision rule for the test is: accept H_0 if the sample mean $\bar{x} \leq 155$
reject H_0 if $\bar{x} > 155$.
- a** Find $P(\text{Type I error})$.
 - b** Suppose the true mean is $\mu = 159$.
 - i** Find the probability of a Type II error.
 - ii** Suppose a new critical value is used for the decision rule. Determine the critical value for $\bar{x}$ so that $P(\text{Type I error}) = P(\text{Type II error})$.
- 3** A sample of size 30 is taken from a normal population with unknown mean μ and known variance 7.5. The sample is used to test the hypotheses $H_0: \mu = 37$, $H_1: \mu < 37$.
- a** Determine the decision rule, in terms of the sample mean $\bar{x}$, for the test so that $P(\text{Type I error}) = 0.05$.
 - b** For the case where the true mean is $\mu = 36$, find the power of the test.
 - c** Find the true value of μ given that $P(\text{Type I error}) = 0.05$ and $P(\text{Type II error}) = 0.1$ for this test.
- 4** The weight of a pumpkin from a very large crop is normally distributed with unknown mean μ and standard deviation 0.7 kg.
Using a random sample of 15 pumpkins from the crop, a statistician conducts a hypothesis test at the 5% level of significance.
Suppose that the true value of the mean is actually 6.4 kg.
- a** For the test $\mu > 6 \text{ kg}$, calculate the probability of a Type II error.
 - b** For the test $\mu \neq 6 \text{ kg}$, calculate the power of the test.
 - c** For the test $\mu < 6 \text{ kg}$, calculate the power of the test.
- 5** The length of a beam produced in a manufacturing process is normally distributed with standard deviation 0.15 m. The manufacturer claims the mean length of such beams is 3.5 m.
A random sample of 20 beams is taken and their mean length $\bar{x}$ is calculated. The value $\bar{x}$ is used to test the manufacturer's claim.
- a** State suitable hypotheses for a two-tailed test.
 - b** For a level of significance of 1%, define the critical region for $\bar{x}$.
 - c** Calculate the probability of making a Type II error if the true mean length is 3.4 m.

ERROR ANALYSIS WITH OTHER DISTRIBUTIONS

Consider a hypothesis test where the null hypothesis H_0 and alternative hypothesis H_1 are defined and the decision rule for accepting or rejecting H_0 is given either numerically or in abstract terms. We may be required to use any one of our known probability distributions to calculate:

$$P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0 \text{ is true}) = \alpha = \text{level of significance for the test.}$$

$$P(\text{Type II error}) = P(\text{Accept } H_0 \mid H_1 \text{ is true}) = 1 - (\text{power of the test}).$$

There are many possible questions of this form. We provide here only a few examples and exercises to demonstrate the general approach to such problems.

The following results may be useful:

Suppose $X_1, X_2, X_3, \dots, X_n$ are independent discrete random variables, and that $S_n = X_1 + X_2 + X_3 + \dots + X_n$.

- If X_i are Bernoulli random variables $X_i \sim B(1, p)$, then $S_n \sim B(n, p)$.
- If X_i are Poisson random variables $X_i \sim \text{Po}(\lambda_i)$, then $S_n \sim \text{Po}(\lambda_1 + \lambda_2 + \lambda_3 + \dots + \lambda_n)$.
- If X_i are Geometric random variables $X_i \sim \text{Geo}(p)$, then $S_n \sim \text{NB}(n, p)$.
- If X_i are Negative binomial random variables $X_i \sim \text{NB}(r_i, p)$, then $S_n \sim \text{NB}(r_1 + r_2 + r_3 + \dots + r_n, p)$.

Example 56

To test whether a coin is fair, the following decision rule is adopted:

Toss the coin 180 times. If the number of heads obtained X is between 75 and 105 inclusive, accept the hypothesis that the coin is fair, otherwise reject it.

- a Define the null hypothesis H_0 and alternative hypothesis H_1 .
- b
 - i Define a Type I error.
 - ii Find the probability of making a Type I error.
 - iii What is the level of significance for this test?
- c Suppose the decision rule is changed so that $P(\text{Type I error}) \approx 0.05$. What is the new decision rule?
- d The coin is actually biased, and the probability of obtaining a head with each toss is 0.65. Using the original decision rule, find $P(\text{Type II error})$.

- a H_0 : The coin is fair, so $p = \text{probability of obtaining a head in one coin toss} = 0.5$

H_1 : The coin is biased, so $p \neq 0.5$.

- b
 - i A Type I error is rejecting H_0 when H_0 is true. This means deciding the coin is biased when it is in fact fair.
 - ii

$$\begin{aligned}
 &P(\text{Type I error}) \\
 &= P(\text{Reject } H_0 \mid H_0 \text{ is true}) \\
 &= P(X \leq 74 \text{ or } X \geq 106 \mid p = 0.5) \quad \{\text{where } X \sim B(180, 0.5)\} \\
 &= 1 - P(75 \leq X \leq 105) \\
 &= 1 - [P(X \leq 105) - P(X \leq 74)] \\
 &\approx 0.0206
 \end{aligned}$$

iii The test is at about the 2% level of significance.

- c Assuming H_0 is true, $X \sim B(180, 0.5)$, and for the hypotheses in a we have a two-tailed test, where the total area of the critical region is 0.05.

Solving $P(X \leq k) = 0.025$ gives $k = 77$

Solving $P(X \leq k) = 0.975$ gives $k = 103$

Check: $1 - P(77 \leq X \leq 103)$
 $= 1 - (P(X \leq 103) - P(X \leq 76))$
 ≈ 0.0439

For this discrete distribution, this is as close as we can get to 0.05 without exceeding it.



The new decision rule is:

Toss the coin 180 times. If the number of heads obtained X satisfies $77 \leq X \leq 103$, accept the null hypothesis that the coin is fair, otherwise reject it.

- d If $p = 0.65$, then $X \sim B(180, 0.65)$.
 $P(\text{Type II error}) = P(75 \leq X \leq 105 \mid X \sim B(180, 0.65))$
 ≈ 0.0374

EXERCISE H.6

- 1 A tetrahedral die has faces marked 1, 2, 3, and 4. To test whether the die is fair for rolling a 4, the following decision rule is adopted:

Roll the die 300 times. If the number of 4s obtained, X , is between 62 and 88 inclusive, we accept the hypothesis that the die is fair for rolling a 4. Otherwise, we reject it.

 - a Define the null hypothesis H_0 and the alternative hypothesis H_1 .
 - b
 - i Define a Type I error for this example.
 - ii Find the probability of making a Type I error.
 - iii What is the level of significance for the test?
 - c Suppose the decision rule is changed so that $P(\text{Type I error}) \approx 0.02$. What is the new decision rule?
 - d The die is actually biased with $P(\text{rolling a 4}) = 0.32$. Use the original decision rule to find $P(\text{Type II error})$.
- 2 A machine fills each can of fizzy drink with volume $Y \text{ cm}^3$, where Y is normally distributed with mean μ and standard deviation 2 cm^3 .

The mean μ is believed to be 330 cm^3 . In order to check this value, a random sample of 16 cans is selected, and the sample mean $\bar{y}$ is calculated.

The following hypotheses are set up: $H_0: \mu = 330$
 $H_1: \mu \neq 330$.

The critical region is defined as $\{\bar{y} < 329\} \cup \{\bar{y} > 331\}$.

 - a Find the significance level for this test.
 - b If the true value of μ is found to be 328 cm^3 , find the probability of a Type II error with this test.

Example 57

A random variable X representing the number of successes in 270 trials can be modelled by a binomial distribution with parameters $n = 270$ and p , whose value is unknown. A significance test is performed, based on a sample value of x_0 , to test the null hypothesis $p = 0.6$ against the alternative hypothesis $p > 0.6$. The probability of making a Type I error is 0.05.

- a Find the critical region for x_0 .
- b Find the probability of making a Type II error in the case when p is actually 0.675.

- a $H_0: p = 0.6, H_1: p > 0.6$.

Since $n = 270$ is very large, the binomial random variable X can be approximated by a normal random variable $X_c \sim N(np, np(1-p))$

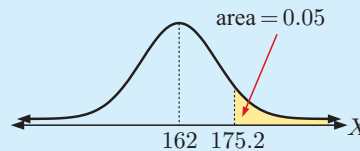
$$\begin{aligned} \text{where } np &= 270 \times 0.6 & \text{and } np(1-p) \\ &= 162 & = 270 \times 0.6 \times 0.4 \\ & & = 64.8 \end{aligned}$$

$$\therefore X_c \sim N(162, 64.8)$$

The critical region is $X_c > 175.2$

But X is discrete,

$\therefore$ the critical region is $X \geq 176$.



- b If $p = 0.675$, $np = 270 \times 0.675$ and $np(1-p)$
- $$\begin{aligned} &= 182.25 & = 270 \times 0.675 \times 0.325 \\ & & \approx 59.231 \end{aligned}$$

Now $X \sim B(270, 0.675)$ can be approximated by $X_c \sim N(182.25, 59.231)$, and from a the critical region is $X > 175.2$

$$\begin{aligned} \therefore \text{P(Type II error)} &= P(H_0 \text{ is accepted} \mid H_1 \text{ is true}) \\ &= P(X \leq 175.2 \mid p = 0.675) \\ &\approx 0.180 \text{ or } 18.0\%. \end{aligned}$$

Example 58

A discrete random variable X has a Poisson distribution with unknown mean m . We wish to test the hypothesis $H_0: m = 2$ against $H_1: m \neq 2$.

A random sample $\{x_1, \dots, x_{12}\}$ of 12 independent values is taken from X , with replacement.

The decision rule is: accept H_0 if $16 \leq \sum_{i=1}^{12} x_i \leq 33$, otherwise reject it.

- a Define the critical region for $S = \sum_{i=1}^{12} x_i$.
- b
 - i Define a Type I error.
 - ii Calculate $P(\text{Type I error})$.
 - c The true value of $m = 2.5$.
 - i Define a Type II error.
 - ii Calculate $P(\text{Type II error})$.

- a Since X is a discrete random variable with values $0, 1, 2, \dots$, so is S .

$\therefore$ the critical region is $\{0 \leq S \leq 15\} \cup \{S \geq 34\}$.

- b i A Type I error is rejecting H_0 when H_0 is in fact true. This means deciding $m \neq 2$ when in fact $m = 2$ and $X \sim \text{Po}(2)$.

- ii Assuming H_0 is true, $X \sim \text{Po}(2)$ and $\therefore S \sim \text{Po}(24)$.

$$\begin{aligned} P(\text{Type I error}) &= P(S \text{ is in the critical region} \mid S \sim \text{Po}(24)) \\ &= P(0 \leq S \leq 15 \text{ or } S \geq 34) \\ &= P(S \leq 15) + 1 - P(S \leq 33) \\ &\approx 0.03440 + 1 - 0.9686 \\ &\approx 0.0658 \end{aligned}$$

- c i A Type II error is accepting H_0 when H_0 is in fact false. This means accepting $m = 2$ when in fact $m = 2.5$.
- ii If $X \sim \text{Po}(2.5)$, then $S \sim \text{Po}(30)$.
- $$\begin{aligned} \therefore P(\text{Type II error}) &= P(16 \leq S \leq 33) \\ &= P(S \leq 33) - P(S \leq 15) \\ &\approx 0.743 \end{aligned}$$

- 3 Eri has a coin which has probability p of giving a head when tossed. She believes that the coin is fair. However, Eri's friend, Mayuki, thinks that the coin is biased, and that $p > 0.5$. To determine which of them is correct, Eri tosses the coin 12 times. Let X denote the number of heads obtained.

- a State appropriate null and alternative hypotheses.
- b Eri rejects the null hypothesis if $X \geq 10$.
- i What name is given to the region $X \geq 10$?
- ii What is meant by the significance level of a hypothesis test? Find its value for this test.
- c In fact, the coin is biased and $p = 0.6$. Find the probability of a Type II error.
- d If Eri uses the decision rule from b, what type of error is Eri at risk of making?

- 4 To find the age of an archaeological specimen, a researcher measures the emission of radioactive particles. The number of particles X emitted in n minutes is said to have a Poisson distribution with parameter $n\lambda$, where the value of λ depends upon the age of the specimen.

Two hypotheses concerning the age of one particular specimen are put forward:

H_0 : the specimen is 5000 years old, in which case $\lambda = 2$

H_1 : the specimen is 10000 years old, in which case $\lambda = 5$.

It is decided to count the number of radioactive particles X emitted in n minutes, and accept H_0 if $X \leq 3$, or reject H_0 if $X \geq 4$.



- a Suppose $n = 1$. Find the probability of:
- i rejecting H_0 given that H_0 is true
- ii accepting H_0 given that H_1 is true.
- b In order that the researcher can publish his findings, the probability of accepting H_0 given that H_1 is true, must be less than 0.005.
- i Show that the minimum number of complete minutes for which counting should be recorded is three.
- ii Find the corresponding probability of rejecting H_0 given that H_0 is true in this case.

Example 59

A discrete random variable X has a geometric distribution with unknown parameter p . We wish to test the hypothesis $H_0: p = 0.3$ against $H_1: p \neq 0.3$.

A random sample $\{x_1, x_2, \dots, x_{10}\}$ of independent values is taken from X , with replacement.

The critical region is defined as $\left\{ \sum_{i=1}^{10} x_i \leq 18 \right\} \cup \left\{ \sum_{i=1}^{10} x_i \geq 55 \right\}$.

- a For which values of $S = \sum_{i=1}^{10} x_i$ will the null hypothesis be accepted?
- b
 - i State the distribution of S , under the null hypothesis.
 - ii Define a Type I error.
 - iii Calculate the level of significance for this test.
- c The true value of p is $p = 0.2$.
 - i Define a Type II error.
 - ii Calculate the power of the test.

- a X is a discrete random variable, and $\therefore$ so is S .

Hence the null hypothesis will be accepted for values of S such that $19 \leq S \leq 54$.

- b i If $X \sim \text{Geo}(0.3)$ then $S \sim \text{NB}(10, 0.3)$ has a negative binomial distribution.
 ii A Type I error is rejecting H_0 when H_0 is in fact true. This means deciding $p \neq 0.3$ when in fact $p = 0.3$.

- iii The level of significance

= P(Type I error)

= $P(S \leq 18 \text{ or } S \geq 55 \mid H_0 \text{ is true})$

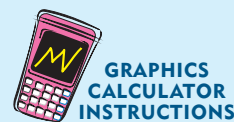
= $P(S \leq 18 \text{ or } S \geq 55 \mid S \sim \text{NB}(10, 0.3))$

= $P(S \leq 18) + (1 - P(S \leq 54))$

= $\sum_{i=10}^{18} \left(\binom{i-1}{9} (0.3)^{10} (0.7)^{i-10} \right) + \left(1 - \sum_{i=10}^{54} \left(\binom{i-1}{9} (0.3)^{10} (0.7)^{i-10} \right) \right)$

≈ 0.04026 {using technology}

$\approx 4\%$



- c i A Type II error is accepting H_0 when H_0 is false. This means accepting $p = 0.3$ when in fact $p = 0.2$.

- ii Power of the test

= $1 - \text{P}(\text{Type II error})$

= $1 - P(19 \leq S \leq 54 \mid p = 0.2)$

= $1 - P(19 \leq S \leq 54 \mid S \sim \text{NB}(10, 0.2))$

= $1 - (P(S \leq 54) - P(S \leq 18))$ when $S \sim \text{NB}(10, 0.2)$

= $1 - \left\{ \sum_{i=10}^{54} \left(\binom{i-1}{9} (0.2)^{10} (0.8)^{i-10} \right) - \sum_{i=10}^{18} \left(\binom{i-1}{9} (0.2)^{10} (0.8)^{i-10} \right) \right\}$

$\approx 1 - \{0.66039 - 0.0009109\}$

≈ 0.341

$\approx 34\%$

- 5** A box is known to contain either 20 white counters and 80 black counters, H_0 , or 50 white counters and 50 black counters, H_1 . In order to test hypothesis H_0 against H_1 , four counters are drawn at random from the box, without replacement. If all four counters are black H_0 is accepted; otherwise it is rejected.
- Find the probabilities of Type I and Type II errors for this test.
 - Determine whether the decision rule “If either three or four counters drawn are black, H_0 is accepted; otherwise it is rejected” gives a test with more power.
- 6** A magician claims that he can roll a six with a fair die on average nine times out of ten.
- Calculate the probability that he will roll five or more sixes in six rolls, assuming:
 - his claim is true
 - he can roll a six, on average, only once in every six rolls.
 - To test the magician’s claim, he is invited to roll the die six times, his claim being accepted if he rolls at least four sixes. Find the probability that the test will:
 - accept the magician’s claim when hypothesis **a ii** is true
 - reject the claim when it is justified, that is, when hypothesis **a i** is true.
- 7** A random variable X has a Poisson distribution with mean m , where m equals 3 or 4. To test the value of m the following hypotheses are defined:
- $$H_0: m = 3$$
- $$H_1: m = 4.$$
- A random sample $\{x_1, x_2, \dots, x_9\}$ of independent values is taken from X , with replacement. If $\sum_{i=1}^9 x_i \leq 37$, then H_0 is accepted, otherwise H_1 is accepted.
- Find the level of significance for this test.
 - Calculate the power of the test.
- 8** A random variable X is known to have a geometric distribution with parameter p which is either 0.25 or 0.38. To test the value of p , two hypotheses are defined:
- $$H_0: p = 0.25$$
- $$H_1: p = 0.38.$$
- A random sample $\{x_1, x_2, \dots, x_{12}\}$ of independent values is taken from X , with replacement. Let $S = \sum_{i=1}^{12} x_i$. If $30 \leq S \leq 71$, then H_0 is accepted, otherwise H_1 is accepted.
- Define the critical region in terms of S .
 - Find $P(\text{Type I error})$.
 - Find $P(\text{Type II error})$.

I BIVARIATE STATISTICS

In this section we consider two variables X and Y which are counted or measured on the same individuals from a population. Since two variables are being considered, this is **bivariate statistics**.

For example, for the given population, variables X and Y could be:

<i>Population</i>	<i>X</i>	<i>Y</i>
A class of students	A student's mark in Physics	A student's mark in Mathematics
A collection of plants of the same species	The length of the stem of the plant	The distance of the plant from a water source
A class of students	A student's mark in Drama	A student's mark in Mathematics

We are interested in the following questions:

- Are X and Y **dependent** or **independent**?
- If X and Y are dependent, what is the nature of the relationship between X and Y ?
- Is the relationship between X and Y **linear**? If so, we can write $Y = a + bX$ or $X = c + dY$ for some constants a, b, c, d .
- If the relationship between X and Y is approximately linear, what is the **strength** of the linear dependence?
- Can we construct a reliable **linear model** $Y = a + bX$ (or $X = c + dY$) and use it to **predict** a value of Y given a value of X , or a value of X given a value of Y ?

CORRELATION AND CAUSATION

Intuitively, **correlation** between two variables X and Y implies some sort of dependence. As one variable changes, the other variable also changes. However, correlation does not imply **causation**. Although two variables might increase or decrease in a related way, it is not necessary that a change in one variable *causes* a change in the other variable. It is possible there is something else influencing both X and Y , or it may simply be coincidence.

For example, for a given population, it is observed over a period of time that:

- the number X of violent computer games sold per year increases
- the number Y of juvenile criminal convictions per year increases
- the number W of cupcakes sold per year increases.

We may find a correlation between X and Y , and we may see a link whereby a change in X might *cause* a change in Y . However, proving causation requires much further analysis.

In contrast, it would appear unreasonable to suggest that a change in X might cause a change in W .

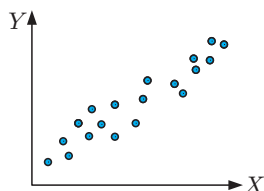
SCATTER DIAGRAMS

Consider two variables X and Y measured on the same population. Each of X and Y has a distribution of values.

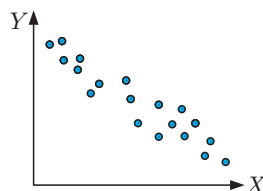
If there are n individuals being considered, we let x_i and y_i be the values of X and Y respectively measured on the i th individual.

The n pairs $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ form the **joint probability distribution** of X and Y . The graph of the points $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ is called a **scatter diagram** (or **scatter plot**).

The scatter diagram may be used to judge whether X and Y are independent, or whether there is a linear relationship between them. If all the points in the scatter diagram lie near or on a straight line, we say there is a **linear correlation** between X and Y .



or

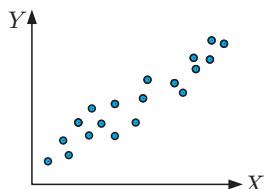


If Y tends to increase as X increases, we have a **positive linear correlation**.

For example:

X = a student's mark in Physics

Y = a student's mark in Mathematics.

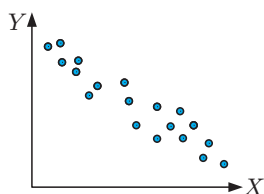


If Y tends to decrease as X increases, we have a **negative linear correlation**.

For example:

X = the length of the stem of a plant

Y = the distance of the plant from a water source.

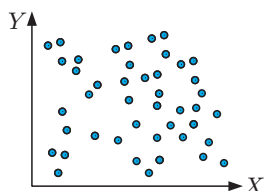


If there is no relationship between X and Y (linear or otherwise) then there is **no correlation** between X and Y . We say that they are **uncorrelated**. X and Y are likely to be independent.

For example:

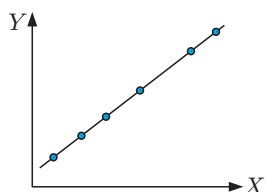
X = a student's mark in Drama

Y = a student's mark in Mathematics.

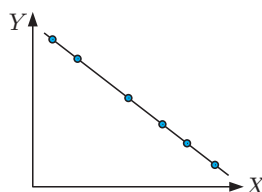


If all points (x, y) lie on a straight line, then we have a **perfect linear correlation**.

positive



negative



THE (SAMPLE) PRODUCT MOMENT CORRELATION COEFFICIENT R AND ITS OBSERVED VALUE r

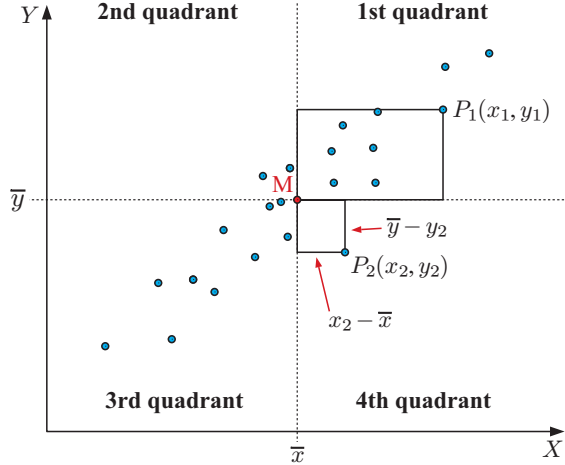
Consider a scatter diagram of the joint probability distribution (X, Y) which includes a random sample of n paired values $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$.

Let $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ be the mean of the y -values.

Let $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ be the mean of the x -values.

The point $M(\bar{x}, \bar{y})$ is called the **mean point**.

The lines $Y = \bar{y}$ and $X = \bar{x}$ are called the **mean lines**. They divide the scatter diagram into four quadrants.



For a point $P(x, y)$, consider the quantity $(x - \bar{x})(y - \bar{y})$, equal to the product of the signed deviations of the point from each of the mean lines.

- If (x, y) lies in the 1st or 3rd quadrants, then $(x - \bar{x})(y - \bar{y}) > 0$.
- If (x, y) lies in the 2nd or 4th quadrants, then $(x - \bar{x})(y - \bar{y}) < 0$.

For the n independent pairwise measured values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, of X and Y on a random sample of n individuals from a population, the **observed value of the (sample) product moment correlation coefficient R** , is

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad \text{where } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Since there are many possible samples in a population, we note that x_i has distribution X_i identical to X , and y_i has distribution Y_i identical to Y , for $i = 1, \dots, n$.

For any such sample, we define the **(sample) product moment correlation coefficient** to be:

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad \text{where } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ and } \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

We can also write $r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_X} \right) \left(\frac{y_i - \bar{y}}{\sigma_Y} \right) = \frac{1}{n} \sum_{i=1}^n (z_{x_i})(z_{y_i})$

where $z_{x_i} = \frac{x_i - \bar{x}}{\sigma_X}$ is the standardised score of x_i , using $\sigma_X = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$,

and $z_{y_i} = \frac{y_i - \bar{y}}{\sigma_Y}$ is the standardised score of y_i , using $\sigma_Y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}$.

We see that r is the average of the product of the standardised signed deviations of point (x_i, y_i) from the mean lines. Therefore:

- $r > 0$ implies most points lie in the 1st and 3rd quadrants of the scatter diagram. This corresponds to a positive linear correlation.
- $r < 0$ implies most points lie in the 2nd and 4th quadrants of the scatter diagram. This corresponds to a negative linear correlation.
- $r = 0$ implies the points are spread symmetrically amongst all four quadrants. This corresponds to the variables being **uncorrelated** (in the linear sense).

Example 60

For each data set:

- Draw a scatter diagram including the mean point $M(\bar{x}, \bar{y})$ and mean lines $Y = \bar{y}$ and $X = \bar{x}$.
- Discuss whether there is a linear correlation between X and Y .
- Calculate the standard deviations σ_X and σ_Y .
- Calculate r .

a

X	2	4	6	8	10	12	14
Y	35	30	25	20	15	10	5

b

X	1	4	7	10	13	16	19
Y	5	9	13	17	21	25	29

c

X	2	3	3	3	3	4	5	5	5	5	6
Y	3	1	2	4	5	3	1	2	4	5	3

d

X	2	4	6	8	10	12	14
Y	6	16	22	24	22	16	6

- a**
- $\bar{x} = 8, \bar{y} = 20$
 $\therefore M(\bar{x}, \bar{y}) = (8, 20)$
 The mean lines are $X = \bar{x} = 8$ and $Y = \bar{y} = 20$.

- There is a perfect negative correlation between X and Y .

- $\sigma_X = 4, \sigma_Y = 10$

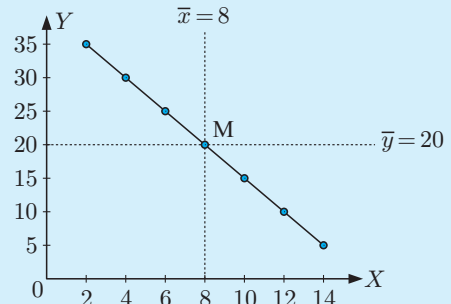
iv

$$r = \frac{1}{7} \sum_{i=1}^7 \left(\frac{x_i - \bar{x}}{\sigma_X} \right) \left(\frac{y_i - \bar{y}}{\sigma_Y} \right)$$

$$= \frac{1}{7} \left\{ \left(\frac{2-8}{4} \right) \left(\frac{35-20}{10} \right) + \left(\frac{4-8}{4} \right) \left(\frac{30-20}{10} \right) + \left(\frac{6-8}{4} \right) \left(\frac{25-20}{10} \right) + \left(\frac{8-8}{4} \right) \left(\frac{20-20}{10} \right) \right.$$

$$\left. + \left(\frac{10-8}{4} \right) \left(\frac{15-20}{10} \right) + \left(\frac{12-8}{4} \right) \left(\frac{10-20}{10} \right) + \left(\frac{14-8}{4} \right) \left(\frac{5-20}{10} \right) \right\}$$

$$= -1$$

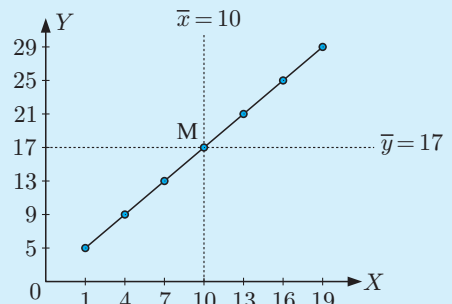


- b**
- $\bar{x} = 10, \bar{y} = 17$
 $\therefore M(\bar{x}, \bar{y}) = (10, 17)$
 The mean lines are $X = \bar{x} = 10$ and $Y = \bar{y} = 17$.

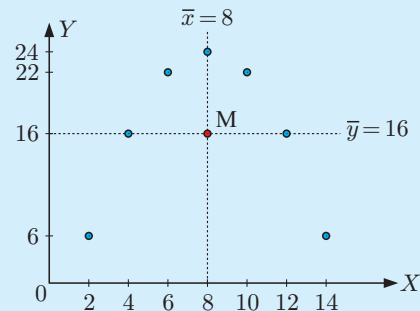
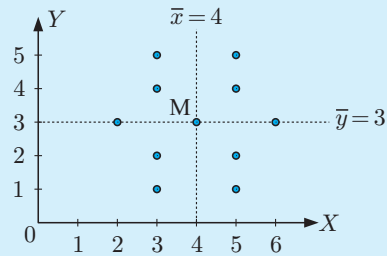
- There is a perfect positive correlation between X and Y .

- $\sigma_X = 6, \sigma_Y = 8$

- iv** Using a similar calculation to **a iv**, $r = 1$.



- c** **i** $\bar{x} = 4$, $\bar{y} = 3$
 $\therefore M(\bar{x}, \bar{y}) = (4, 3)$
 The mean lines are $X = \bar{x} = 4$ and $Y = \bar{y} = 3$.
- ii** There is no linear correlation between X and Y . The variables are possibly independent.
- iii** $\sigma_X \approx 1.206$, $\sigma_Y \approx 1.348$
- iv** $r = 0$
- d** **i** $\bar{x} = 8$, $\bar{y} = 16$
 $\therefore M(\bar{x}, \bar{y}) = (8, 16)$
 The mean lines are $X = \bar{x} = 8$ and $Y = \bar{y} = 16$.
- ii** There is a quadratic relationship between the variables, so they are dependent. However, there is no linear correlation between them.
- iii** $\sigma_X \approx 4$, $\sigma_Y \approx 6.928$ **iv** $r = 0$



From the above example we observe:

- perfect positive linear correlation has $r = 1$
- perfect negative linear correlation has $r = -1$
- if there is no linear correlation between X and Y , then $r = 0$.

This example shows the extreme cases, and shortly we will prove that $-1 \leq r \leq 1$.

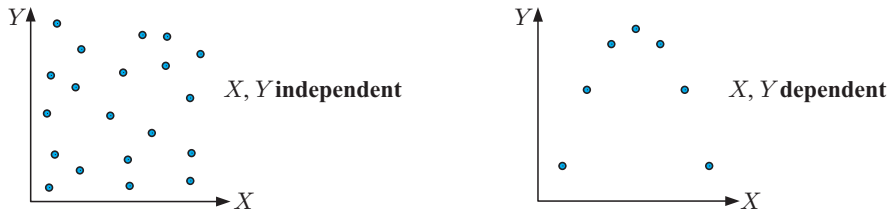
Positive correlation

$r = 1$	perfect positive correlation	
$0.95 \leq r < 1$	very strong positive correlation	
$0.87 \leq r < 0.95$	strong positive correlation	
$0.5 \leq r < 0.87$	moderate positive correlation	
$0 < r < 0.5$	weak positive correlation	

Negative correlation

$r = -1$	perfect negative correlation	
$-1 < r \leq -0.95$	very strong negative correlation	
$-0.95 < r \leq -0.87$	strong negative correlation	
$-0.87 < r \leq -0.5$	moderate negative correlation	
$-0.5 < r < 0$	weak negative correlation	

We note that even for the case of no linear correlation, where $r = 0$, X and Y may be independent or dependent. For example:



EXERCISE 1.1

- 1 a Draw the scatter diagram for the bivariate data set

x	2	2	2	2	3	3	3	4	4	5
y	8	10	12	14	8	10	12	8	10	8

- b Hence determine whether there is positive, negative, or no correlation.
 c Draw the lines $x = \bar{x}$ and $y = \bar{y}$ on the diagram. Do these lines and the four quadrants confirm your decision in b?
 d Use the formula $r = \frac{1}{n} \sum_{i=1}^n (z_{x_i})(z_{y_i})$ to show that $r = -0.5$.

- 2 a Use $r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$ to show that $r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \bar{x}^2\right) \left(\sum_{i=1}^n y_i^2 - n \bar{y}^2\right)}}$.

- b Hence show that $r = -0.5$ for the data set in question 1.

- 3 Consider the following set of bivariate data about the average daily food consumption of 12 obese adults.

Weight (x kg)	96	87	84	75	98	88	93	81	89	64	68	78
Food consumption ($y \times 100$ calories per day)	35	34	35	29	41	35	36	30	34	26	28	31

- a Draw a scatter diagram for the data set.
 b Describe the correlation between x and y .
 c Calculate the product moment correlation coefficient.
 4 A selection of students were asked how many phone calls and text messages they had received the previous day. The results are shown below.

Student	A	B	C	D	E	F	G	H
Phone calls received	4	7	1	0	3	2	2	4
Text messages received	6	9	2	2	5	8	4	7

- a Draw a scatter diagram of the data.
 b Calculate r .
 c Describe the linear correlation between *phone calls received* and *text messages received*.

- 5 A basketballer takes 20 shots from each of ten different positions marked on the court. The table below shows how far each position is from the goal, and how many shots were successful:

Position	A	B	C	D	E	F	G	H	I	J
Distance from goal (x m)	2	5	3.5	6.2	4.5	1.5	7	4.1	3	5.6
Successful shots (y)	17	6	10	5	8	18	6	8	13	9

- Draw a scatter diagram of the data.
- Do you think r will be positive or negative?
- Calculate the value of r .
- Describe the linear correlation between these variables.
- Copy and complete:
As the distance from goal increases, the number of successful shots generally
- Is there a causal relationship between these variables?

- 6 Consider the data set:

x	1	2	3
y	1	3	4

- Calculate r for the data set.
- Let $u = 2x + 1$ and $v = 3x - 4$.
 - List the data (u, v) in a table.
 - Calculate r for this data.
- Let $u = -3x + 5$ and $v = 3x - 4$.
 - List the data (u, v) in a table.
 - Calculate r for this data.
- Let $u = 2x + 1$ and $v = -3y - 1$.
 - List the data (u, v) in a table.
 - Calculate r for this data.
- Let $u = -2x + 1$ and $v = -3y - 1$.
 - List the data (u, v) in a table.
 - Calculate r for this data.
- Compare your answers from **a** to **e**. What do your results suggest?

CORRELATION BETWEEN TWO RANDOM VARIABLES X AND Y AND THE (POPULATION) PRODUCT MOMENT CORRELATION COEFFICIENT ρ

So far we have considered a measure r for the strength of correlation between variables X and Y using observed values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Such a set of values may well be from a population of n individuals. However, in practice it is likely to be only a sample of values from a larger population on which variables X and Y are measured.

We now establish the theory to examine linear correlation between random variables X and Y for a general population.

Consider the distribution of X and the distribution of Y for the whole population.

Let $E(X) = \mu_X$ denote the mean value of X

$E(Y) = \mu_Y$ denote the mean value of Y

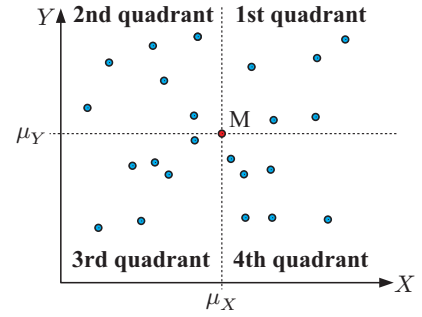
$\text{Var}(X) = \sigma_X^2$ denote the variance of X

$\text{Var}(Y) = \sigma_Y^2$ denote the variance of Y .

Motivated by our previous work, we consider a scatter diagram for pairs (x, y) for X and Y measured on each individual in the population.

Let $M(\mu_X, \mu_Y)$ be the **mean point** and $x = \mu_X$, $y = \mu_Y$ be the **mean lines**.

The mean lines divide the scatter diagram into four quadrants.

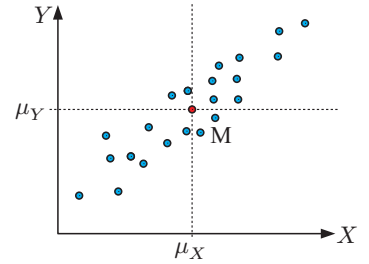


For a pair (x, y) , the quantity $(x - \mu_X)(y - \mu_Y)$ gives a signed measure of the deviations of x and y from their respective means μ_X and μ_Y . The average value of these deviations, $E((X - \mu_X)(Y - \mu_Y))$, provides a measure of the strength of linear dependence between X and Y .

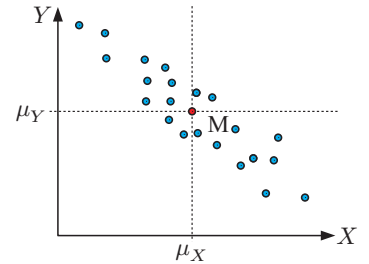
The **covariance** of X and Y is $\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$.

We note that:

- The larger $|\text{Cov}(X, Y)|$ the greater the linear dependence of X and Y .
- If $\text{Cov}(X, Y)$ is large and positive, the majority of points (x, y) lie in the first and third quadrants of the associated scatter diagram. X and Y have a positive linear correlation.



- If $\text{Cov}(X, Y)$ is large and negative, the majority of points (x, y) lie in the second and fourth quadrants of the scatter diagram. X and Y have a negative linear correlation.



Theorem 14

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

Proof:

$$\begin{aligned}
 \text{Cov}(X, Y) &= E((X - \mu_X)(Y - \mu_Y)) \\
 &= E(XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y) \\
 &= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X \mu_Y \\
 &\quad \text{\{by Theorem 7 and since } E(\mu_X \mu_Y) = \mu_X \mu_Y \text{ by Theorem 1\}} \\
 &= E(XY) - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y \\
 &= E(XY) - E(X)E(Y)
 \end{aligned}$$

Theorem 15

If X, Y are **independent** random variables, then $\text{Cov}(X, Y) = 0$.

Proof:

If X, Y are independent, then $E(XY) = E(X)E(Y)$ {**Theorem 7**}

$$\begin{aligned}\therefore \text{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(X)E(Y) - E(X)E(Y) \\ &= 0.\end{aligned}$$

Note that converse of **Theorem 15** is not true. If $\text{Cov}(X, Y) = 0$, it is not necessarily true that X and Y are independent.

We have observed that if $\text{Cov}(X, Y)$ is large and positive, then X and Y have a positive linear correlation. However, since the definition of covariance very much depends on the given scales and values of X and Y , it is difficult to quantify “large”. We therefore standardise the value by defining:

The **(population) product moment correlation coefficient** ρ of random variables X and Y is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

Note that
$$\begin{aligned}\rho &= \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \\ &= E\left(\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right) \\ &= E(Z_X Z_Y)\end{aligned}$$

where $Z_X = \frac{X - \mu_X}{\sigma_X}$, $Z_Y = \frac{Y - \mu_Y}{\sigma_Y}$, and σ_X, σ_Y are the standard deviations of X and Y respectively.

Theorem 16

For X and Y two random variables with finite variances, $\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$ satisfies $-1 \leq \rho \leq 1$.

Proof:

Let $Z_X = \frac{X - \mu_X}{\sigma_X}$, $Z_Y = \frac{Y - \mu_Y}{\sigma_Y}$ where $\text{Var}(X) = \sigma_X^2$ and $\text{Var}(Y) = \sigma_Y^2$.

$$\therefore E(Z_X) = E(Z_Y) = 0.$$

$$\begin{aligned}\text{Also } E(Z_X^2) &= \text{Var}(Z_X) + (E(Z_X))^2 \\ &= 1 + 0^2 \\ &= 1\end{aligned}$$

and similarly $E(Z_Y^2) = 1$.

Let $U = tZ_X + Z_Y$. Since $U^2 \geq 0$, $E(U^2) \geq 0$

$$\begin{aligned}\therefore E((tZ_X + Z_Y)^2) &\geq 0 \\ \therefore E(t^2 Z_X^2 + 2tZ_X Z_Y + Z_Y^2) &\geq 0 \\ \therefore t^2 E(Z_X^2) + 2tE(Z_X Z_Y) + E(Z_Y^2) &\geq 0 \\ \therefore t^2 + 2t\rho + 1 &\geq 0\end{aligned}$$

The discriminant $4\rho^2 - 4$ of this quadratic therefore satisfies $4\rho^2 - 4 \leq 0$
 $\therefore \rho^2 \leq 1$
 $\therefore -1 \leq \rho \leq 1$

It follows that the observed value r of the sample product moment correlation coefficient satisfies $-1 \leq r \leq 1$. This is because, for X with distribution $\{x_1, x_2, \dots, x_n\}$ and Y with distribution $\{y_1, y_2, \dots, y_n\}$, for the n data pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ we have $E(X) = \bar{x}$, $E(Y) = \bar{y}$, $\text{Var}(X) = \sigma_X^2$, $\text{Var}(Y) = \sigma_Y^2$ and therefore $r = \rho$ for this data.

Theorem 17

Let X and Y be random variables with finite variance.

- 1 If X and Y are independent then $\rho = 0$.
- 2 X and Y have a perfect linear correlation if and only if $\rho = \pm 1$.

Proof:

- 1 If X and Y are independent then $\text{Cov}(X, Y) = 0$ {Theorem 15}

$$\therefore \rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = 0.$$

Since X and Y are independent, there is no linear correlation between them.

- 2 ($\Rightarrow$) If $Y = aX + b$, $a, b \in \mathbb{R}$, $a \neq 0$, then $\mu_Y = E(Y) = E(aX + b)$
 $= aE(X) + b$
 $= a\mu_X + b$

$$\text{and } \text{Var}(Y) = a^2\text{Var}(X)$$

$$\begin{aligned} \text{Also } Y - \mu_Y &= aX + b - \mu_Y \\ &= aX + b - (a\mu_X + b) \\ &= a(X - \mu_X) \end{aligned}$$

$$\begin{aligned} \therefore \text{Cov}(X, Y) &= E((X - \mu_X)(Y - \mu_Y)) \\ &= E(a(X - \mu_X)^2) \\ &= aE((X - \mu_X)^2) \\ &= a\text{Var}(X) \end{aligned}$$

$$\therefore \rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{a\text{Var}(X)}{\sqrt{a^2(\text{Var}(X))^2}} = \frac{a}{|a|} = \pm 1$$

($\Leftarrow$) Conversely, suppose $\rho = 1$ or -1 .

$$\begin{aligned} \text{From the proof of Theorem 16 this means } E(U^2) &= t^2 + 2t\rho + 1 \\ &= (t+1)^2 \text{ if } \rho = 1 \text{ or} \\ &= (t-1)^2 \text{ if } \rho = -1. \end{aligned}$$

Case $E(U^2) = (t+1)^2$

Let $t = -1$, so $U = Z_Y - Z_X$.

$$\therefore E(U^2) = 0 \text{ and } E(U) = E(Z_Y - Z_X) = E(Z_Y) - E(Z_X) = 0$$

$$\therefore \text{Var}(U) = E(U^2) - (E(U))^2 = 0 - 0^2 = 0$$

Thus U takes only one value, the mean 0, with probability 1.

$$\therefore 0 = Z_Y - Z_X$$

$$\therefore 0 = \frac{Y - \mu_Y}{\sigma_Y} - \frac{X - \mu_X}{\sigma_X}$$

Rearranging, there exists a linear relationship between X and Y :

$$Y = \frac{\sigma_Y}{\sigma_X} X + (\mu_Y - \sigma_Y \frac{\mu_X}{\sigma_X})$$

Case $E(U^2) = (t - 1)^2$

Let $t = 1$, so $U = Z_X + Z_Y$.

$$\therefore E(U^2) = 0 \quad \text{and} \quad E(U) = E(Z_X + Z_Y) = E(Z_X) + E(Z_Y) = 0$$

$$\therefore \text{Var}(U) = E(U^2) - (E(U))^2 = 0 - 0^2 = 0$$

Thus U takes only one value, the mean 0, with probability 1.

$$\therefore 0 = Z_X + Z_Y$$

$$\therefore 0 = \frac{X - \mu_X}{\sigma_X} + \frac{Y - \mu_Y}{\sigma_Y}$$

Rearranging, there exists a linear relationship between X and Y :

$$Y = -\frac{\sigma_Y}{\sigma_X} X + (\mu_Y + \sigma_Y \frac{\mu_X}{\sigma_X})$$

The (population) product moment correlation coefficient ρ is hence a measure of the strength of the linear correlation between random variables X and Y . Values of ρ near 0 indicate a weak correlation and values of ρ near ± 1 indicate a strong correlation. Values $\rho = \pm 1$ indicate a perfect linear correlation.

In practice, we usually have only a sample from a population, and the observed value r of the (sample) product moment correlation coefficient is used as an estimate for ρ .

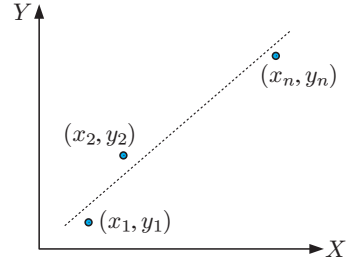
EXERCISE 1.2

- 1 Use the formula given in **Theorem 14** for the covariance $\text{Cov}(X, Y)$ of two random variables X and Y , to show that:
 - a $\text{Cov}(X, X) = \text{Var}(X)$
 - b $\text{Cov}(X, X + Y) = \text{Cov}(X, X) + \text{Cov}(X, Y)$
 - c If $X = c$ a constant, then $\text{Cov}(X, Y) = 0$ for any random variable Y .
- 2 For X and Y two random variables, find $\text{Cov}(X + Y, X - Y)$.
- 3 For X and Y two random variables, show that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.
- 4 Suppose X and Y are two random variables. Show that if $Y = mX + c$ for constants m and c with $m \neq 0$, then $\rho = \pm 1$.
- 5 Let X and Y be two random variables with correlation coefficient ρ .
Let $U = a + bX$ and $V = c + dY$ be two new random variables, for a, b, c, d constants.
Find the product moment correlation coefficient of U and V in terms of ρ .
- 6 If X and Y are random variables with $\sigma_X^2 = 1$, $\sigma_Y^2 = 9$, and $\rho = \frac{1}{9}$, find the exact value of the correlation coefficient between X and $X + Y$.
Hint: Use the result of question 3.

THE TWO LINES OF REGRESSION

Consider the sample data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. When we have a strong linear correlation between X and Y , it is useful to fit a straight line model to the data. If we are given the value of one variable, we can use the model to predict the value of the other.

Conventionally, the “line of best fit” for data is obtained by Gauss’ **method of least squares**, as follows.



THE REGRESSION LINE OF Y ON X

For this regression line we rely more heavily on the measured values of x . We use this line to predict a value of y for a given value of x .

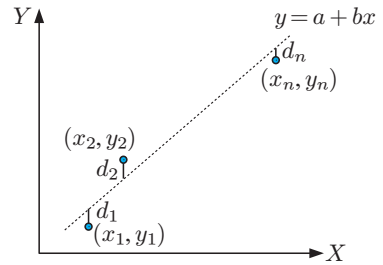
Consider using a line of the form $y = a + bx$ to model the data. The constants a and b are chosen so that the sum of the squared **vertical distances** of points from the line is a minimum. This means we minimise

$$d_1^2 + d_2^2 + \dots + d_n^2 = \sum_{i=1}^n (y_i - a - bx_i)^2.$$

Suppose $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Calculus can be used to show that

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$



The **regression line of Y on X** has equation $(y - \bar{y}) = b(x - \bar{x})$

which is $(y - \bar{y}) = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] (x - \bar{x}).$

The constant b is called the **regression coefficient** for this line.

We use this line of best fit to predict a value of y for a given value of x .

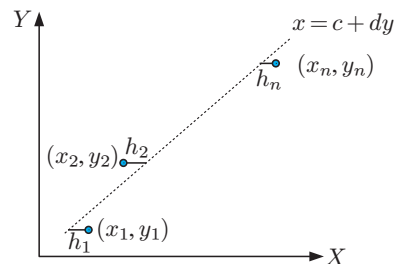


THE REGRESSION LINE OF X ON Y

For this regression line we rely more heavily on the measured values of y . We use this line to predict a value of x for a given value of y .

Consider using a line of the form $x = c + dy$ to model the data. The constants c and d are chosen so that the sum of the squared **horizontal distances** of points from the line is a minimum. This means we minimise

$$h_1^2 + h_2^2 + \dots + h_n^2 = \sum_{i=1}^n (x_i - c - dy_i)^2.$$



Suppose $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Calculus can be used to show that

$$c = \bar{x} - d\bar{y} \quad \text{and} \quad d = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n y_i^2 - n(\bar{y})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

The **regression line of X on Y** has equation $(x - \bar{x}) = d(y - \bar{y})$

which is $(x - \bar{x}) = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} \right] (y - \bar{y}).$

The constant d is called the **regression coefficient** for this line.

We use this line of best fit to predict a value of x for a given value of y .



We notice that the mean point $M(\bar{x}, \bar{y})$ lies on both regression lines. Thus either the two lines are identical or $M(\bar{x}, \bar{y})$ is the unique point of intersection of the two regression lines.

Example 61

Consider the data set from **Exercise I.1** question **3**, about the average daily food consumption of 12 obese adults.

Weight (x kg)	96	87	84	75	98	88	93	81	89	64	68	78
Food consumption ($y \times 100$ calories per day)	35	34	35	29	41	35	36	30	34	26	28	31

- Find the regression line of y on x and the regression line of x on y .
- Find $\bar{x}$ and $\bar{y}$, the mean of each data set.
- Use the appropriate regression line to estimate:
 - the food consumption of an obese adult who weighs 92 kg
 - the weight of an obese adult whose food consumption is 3250 calories per day.

a

	List 1	List 2	List 3	List 4
1	96	35		
2	87	34		
3	84	35		
4	75	29		

1Var XList :List1
1Var Freq :1
2Var XList :List1
2Var YList :List2
2Var Freq :1

1Var XList :List1
1Var Freq :1
2Var XList :List2
2Var YList :List1
2Var Freq :1

LinearReg(ax+b)
a = 0.36418229
b = 2.45446053
r = 0.93452311
r ² = 0.87333345
MSe = 2.40244211
y = ax + b

LinearReg(ax+b)
a = 2.39806678
b = 4.68014059
r = 0.93452311
r ² = 0.87333345
MSe = 15.8195957
y = ax + b

The regression line of y on x is $y = 0.364x + 2.454$.

The regression line of x on y is $x = 2.398y + 4.68$.

- b** The point of intersection of the lines is $M(83.4, 32.8)$. $\therefore \bar{x} \approx 83.4, \bar{y} \approx 32.8$
- c** Using the regression line of y on x with $x = 92$ kg, $y \approx 0.364 \times 92 + 2.454$
 ≈ 35.942

We expect the food consumption will be about 3594.2 calories per day.

- d** Using the regression line of x on y with $y = 32.5$, $x \approx 2.398 \times 32.5 + 4.68$
 ≈ 82.6 kg

We expect the weight will be about 82.6 kg.

Theorem 18

For a random sample of n independent paired values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$:

- 1** The regression line of Y on X and the regression line of X on Y are the same line if and only if $r = \pm 1$, in other words, if and only if X and Y have a perfect linear correlation.
- 2** If $r = 0$ then the regression line Y on X has equation $y = \bar{y}$ (a horizontal line) and the regression line X on Y has equation $x = \bar{x}$ (a vertical line).

Proof:

By definition,
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Define: $s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ to be the **sample covariance**

$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ to be the sample variance of the x -values

$s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$ to be the sample variance of the y -values.

$$\begin{aligned} \therefore r &= \frac{s_{xy}}{s_x s_y} \text{ and the regression line of } Y \text{ on } X \text{ is } (y - \bar{y}) = b(x - \bar{x}) \\ &= \frac{s_{xy}}{s_x^2} (x - \bar{x}) \\ &= \frac{s_{xy}}{s_x} \left(\frac{x - \bar{x}}{s_x} \right) \end{aligned}$$

We multiply by $\frac{1}{s_y}$ to obtain $\left(\frac{y - \bar{y}}{s_y} \right) = r \left(\frac{x - \bar{x}}{s_x} \right) \dots (1)$

Similarly the **regression line of X on Y** is $\left(\frac{x - \bar{x}}{s_x} \right) = r \left(\frac{y - \bar{y}}{s_y} \right) \dots (2)$

- 1** Comparing (1) and (2) we see the two lines are identical if and only if $r = \frac{1}{r}$

$$\therefore r = \pm 1$$

- 2** If $r = 0$, then using equations (1) and (2) we find the regression line of Y on X is $y = \bar{y}$, and the regression line of X on Y is $x = \bar{x}$.

Corollary:

Suppose b is the regression coefficient of the regression line Y on X and d is the regression coefficient of the regression line X on Y .
Then r is the geometric mean of b and d .

Proof:

$$\text{Using the same definitions as above, } bd = \frac{s_{xy}}{s_x^2} \frac{s_{xy}}{s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} = r^2$$

$$r = \sqrt{bd}$$

Theorem 18 (2) verifies the intuitive results that when X and Y are uncorrelated:

- no matter what the value of X , the expected value of Y is $E(Y) = \bar{y}$.
- no matter what the value of Y , the expected value of X is $E(X) = \bar{x}$.

EXERCISE 1.3

- 1 The results of ten people in a Mathematics and a Physics test were:

<i>Mathematics (x)</i>	13	10	8	14	6	11	10	5	12	13
<i>Physics (y)</i>	21	15	14	20	12	16	9	10	17	12

- Determine the product moment correlation coefficient and the equations of the two regression lines.
 - Estimate the Mathematics score of someone who obtained a Physics score of 11.
 - Estimate the Physics score of someone who obtained a Mathematics score of 18.
 - Which of the estimates do you expect to be more accurate? Explain your answer.
- 2 The cholesterol level in the bloodstream (x) and the resting heart beat (y) of 10 people are:

<i>Cholesterol level (x)</i>	5.32	5.54	5.45	5.06	6.13	5.00	4.90	6.00	6.70	4.75
<i>Resting heart beat (y)</i>	55	48	55	53	74	44	49	68	78	51

- Determine the product moment correlation coefficient and the equations of the two regression lines.
 - Estimate the cholesterol level of someone with a resting heart beat of 60.
 - Estimate the resting heart beat of someone with a cholesterol level of 5.8.
- 3 Eight students swim 200 m breaststroke. Their times y in seconds, and arm lengths x in cm, are shown in the table below:

<i>Length of arm (x cm)</i>	78	73	71	68	76	72	63	69
<i>Breaststroke (y seconds)</i>	123.1	123.7	127.3	132.0	120.8	125.0	140.9	129.0

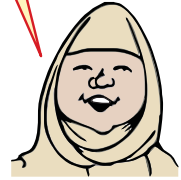
- Determine the product moment correlation coefficient and the equations of the two regression lines.
- Estimate the time to swim 200 m breaststroke for someone with arm length 65 cm.

- 4 Consider the bivariate data given below, the first line of which gives the weight x in kg of 10 men and 9 women and the second line of which gives the total weight y of body fat for each person.

Men	Weight (x kg)	89	88	66	59	93	73	82	77	100	67
	Body fat (y kg)	12	14	9	10	22	13	13	11	19	12
Women	Weight (x kg)	57	68	69	59	62	59	56	66	72	
	Body fat (y kg)	17	22	24	18	18	15	16	22	24	

- Calculate the product moment correlation coefficient for X and Y for:
 - the men
 - the women
 - the 19 people in the data set.
- Determine the percentage of body fat w for each person in the data set.
- Calculate the product moment correlation coefficient for X and W for:
 - the men
 - the women
 - the 19 people in the data set.

You may wish to use a spreadsheet.



THE BIVARIATE NORMAL DISTRIBUTION

For single continuous random variables, the normal distribution is a classic distribution for which we have many theorems and results. In this section we generalise the normal distribution to two continuous random variables X and Y .

X and Y have a **bivariate normal distribution**, or we say X and Y are **jointly normally distributed**, if they have the joint probability density function

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp \left\{ -\frac{\left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right]}{2(1-\rho^2)} \right\}$$

for $-\infty < x < \infty$, $-\infty < y < \infty$, and where

X is normal with mean μ_X and standard deviation σ_X ,

Y is normal with mean μ_Y and standard deviation σ_Y ,

and $\rho = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y}$ is the (population) product moment correlation coefficient of X and Y .

$\exp\{x\} = e^x$



We have seen that if X and Y are any two independent random variables, then by **Theorem 15**, $\text{Cov}(X, Y) = 0$ and therefore $\rho = 0$.

For X and Y having a bivariate normal distribution, the converse of this result is also true.

Theorem 19

For X and Y with a bivariate normal distribution, X and Y are independent if and only if $\rho = 0$.

Proof:

By the remarks preceeding this theorem we need only prove the ($\Leftarrow$) case.

Suppose X and Y have a bivariate normal distribution with correlation coefficient $\rho = 0$.

The joint probability density function of X and Y becomes

$$\begin{aligned} f(x, y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp \left\{ -\frac{1}{2} \left[\left(\frac{x - \mu_X}{\sigma_X} \right)^2 + \left(\frac{y - \mu_Y}{\sigma_Y} \right)^2 \right] \right\} \\ &= \left(\frac{1}{\sigma_X\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu_X}{\sigma_X} \right)^2 \right\} \right) \left(\frac{1}{\sigma_Y\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{y - \mu_Y}{\sigma_Y} \right)^2 \right\} \right) \\ &= f(x)f(y) \end{aligned}$$

where $f(x)$ is the PDF of a normal random variable X with mean μ_X and standard deviation σ_X and $f(y)$ is the PDF of a normal random variable Y with mean μ_Y and standard deviation σ_Y .

Hence any probability calculated in the joint distribution will have form

$$\begin{aligned} P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) &= \int_{y_1}^{y_2} \left(\int_{x_1}^{x_2} f(x, y) dx \right) dy \\ &= \int_{y_1}^{y_2} \left(\int_{x_1}^{x_2} f(x)f(y) dx \right) dy \\ &= \left(\int_{x_1}^{x_2} f(x) dx \right) \left(\int_{y_1}^{y_2} f(y) dy \right) \\ &= P(x_1 \leq X \leq x_2) \times P(y_1 \leq Y \leq y_2) \end{aligned}$$

Hence X and Y are independent.

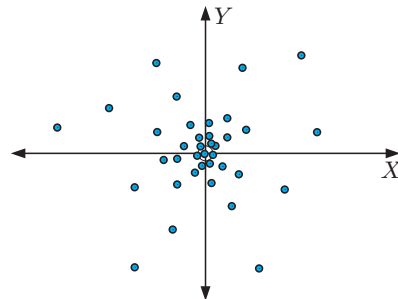
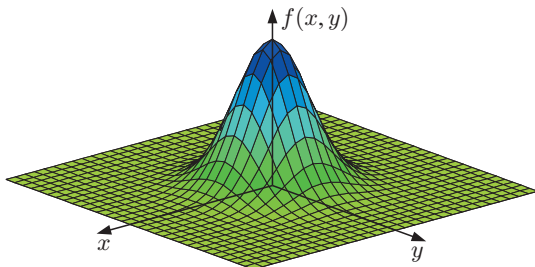
SHAPE OF THE BIVARIATE NORMAL DISTRIBUTION

The bivariate normal distribution takes the shape of a 3-dimensional bell-shaped surface.

Consider the following cases which describe how the different parameters affect the position, shape, and orientation of the surface. Alongside each is a typical scatterplot for X and Y that might result from that distribution.

Case 1:

- $\mu_X = \mu_Y = 0 \Rightarrow$ The peak is above the origin.
- $\sigma_X = \sigma_Y \Rightarrow$ The surface is axially symmetrical.
- $\rho = 0 \Rightarrow$ X and Y are independent.

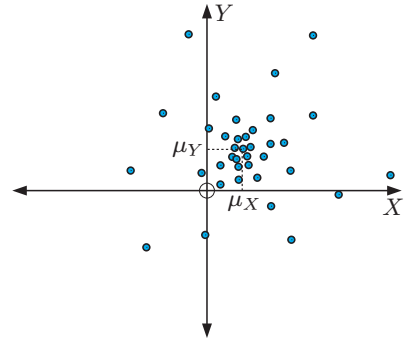
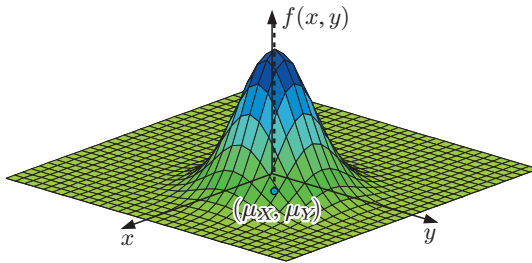


Case 2:

$\mu_X \neq 0, \mu_Y \neq 0 \Rightarrow$ The peak is at $M(\mu_X, \mu_Y)$.

$$\sigma_X = \sigma_Y$$

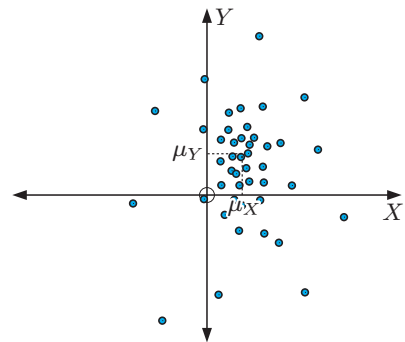
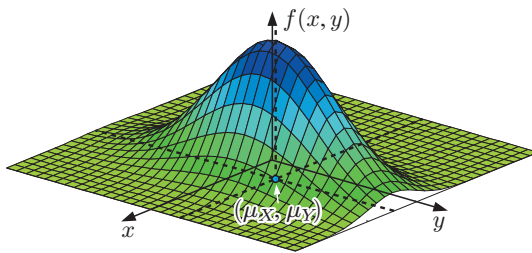
$$\rho = 0$$

**Case 3:**

$$\mu_X, \mu_Y$$

$\sigma_X \neq \sigma_Y \Rightarrow$ The surface has lines of symmetry parallel to the X and Y axes.

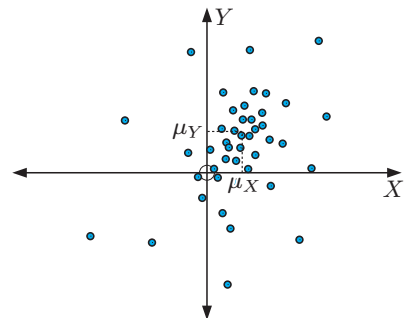
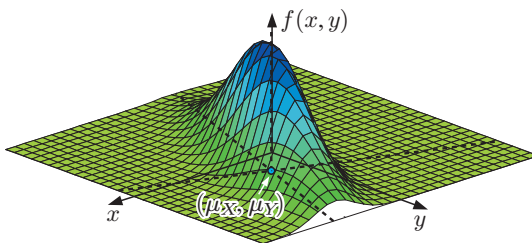
$$\rho = 0$$

**Case 4:**

$$\mu_X, \mu_Y$$

$$\sigma_X \neq \sigma_Y$$

$\rho \neq 0 \Rightarrow$ X and Y are now dependent.
The surface has lines of symmetry *not* parallel to the X and Y axes.



3-D BELL-SHAPED SURFACE



The result in **Case 4** is a probability distribution which favours the points on the scatterplot for X and Y showing linear correlation. For a sufficiently large sample of points, the regression line will approximate one of the lines of symmetry of the bell-shaped probability distribution surface.

HYPOTHESIS TESTING FOR DEPENDENCE OF X AND Y

We have seen that to analyse the strength of linear correlation between two random variables X and Y we use observed values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, and either a scatter diagram or the value of r .

In the case of X and Y having a bivariate normal distribution, a single such value of r can be used in a hypothesis test to determine (at the given level of significance) whether or not $\rho = 0$ for X and Y , and therefore whether or not X and Y are linearly correlated. By **Theorem 19**, this will also tell us whether or not X and Y are independent.

We require the following result:

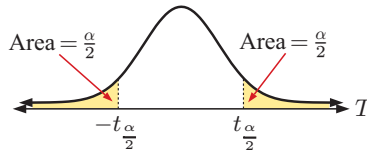
If X and Y have a bivariate normal distribution such that $\rho = 0$, then the sampling distribution $R\sqrt{\frac{n-2}{1-R^2}}$ has the Student's t -distribution with $(n-2)$ degrees of freedom.

Hence for any random sample of n independent paired data values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ from the jointly normal distribution (X, Y) with $\rho = 0$, the observed value r of the sample product moment correlation coefficient is such that $r\sqrt{\frac{n-2}{1-r^2}}$ lies in the $t(n-2)$ -distribution.

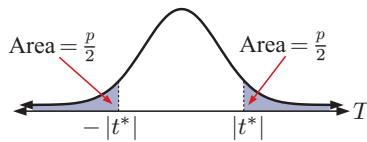
HYPOTHESIS TEST WITH $H_0: \rho = 0$ AND $H_1: \rho \neq 0$

Given a random sample of n independent paired data values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ from a bivariate normal distribution (X, Y) :

- (1) Calculate r , the observed value of the sample product moment correlation coefficient.
- (2) Calculate the test statistic $t^* = r\sqrt{\frac{n-2}{1-r^2}}$.
- (3) **Either** For level of significance α , calculate the **critical values** $\pm t_{\frac{\alpha}{2}}$ for the $t(n-2)$ -distribution which define the two-tailed **critical region**.



or Calculate the p -value $P(T \leq -|t^*|) + P(T \geq |t^*|)$
 $= 2P(T \geq |t^*|)$



- (4) We do not reject H_0 if $-t_{\frac{\alpha}{2}} \leq t^* \leq t_{\frac{\alpha}{2}}$, or if the p -value is not less than α .

We reject H_0 in favour of H_1 if t^* lies in the **critical region** $t^* < -t_{\frac{\alpha}{2}}$ or $t^* > t_{\frac{\alpha}{2}}$, or if the p -value is less than α .

If the result of the hypothesis test is to reject $H_0: \rho = 0$ in favour of $H_1: \rho \neq 0$, we conclude that the bivariate data is correlated. The strength of this correlation is measured by the value of r . In this case the regression line of Y on X and the regression line of X on Y can be calculated, and predictions for values of X and Y can be made.

Example 62

Consider the following data set sampled from a bivariate normal distribution.

x	4	3	4	3	-4	-3	-4	-3
y	3	4	-3	-4	3	4	-3	-4

Conduct hypothesis tests at the 5% and 1% levels of significance, to determine whether or not the two sets of data are correlated.

We test at the 5% and 1% levels: $H_0: \rho = 0$ against $H_1: \rho \neq 0$.

$n = 8$, so there are 6 degrees of freedom.

(1) r is calculated and is found to be zero.

$$(2) \quad t^* = r \sqrt{\frac{n-2}{1-r^2}} = (0) \sqrt{\frac{8-2}{1-(0)^2}} = 0.$$

(3) **Either** Since $t^* = 0$, $-t_{\frac{\alpha}{2}} \leq t^* \leq t_{\frac{\alpha}{2}}$ no matter what the level of significance is.

or The p -value equals 1.

(4) We retain H_0 at both 5% and 1% levels, and conclude X and Y are uncorrelated.

Example 63

The following data set is sampled from a bivariate normal distribution:

x	2	2	2	2	3	3	3	4	4	5
y	8	10	12	14	8	10	12	8	10	8

Conduct a hypothesis test at the 5% level of significance to determine whether or not the two sets of data are correlated.

$H_0: \rho = 0$ $H_1: \rho \neq 0$

(1) $r = -0.5$

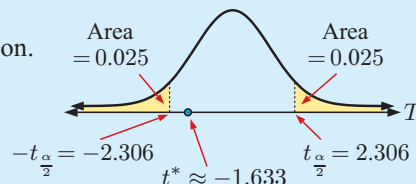
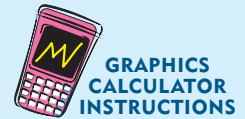
$$(2) \quad t^* = r \sqrt{\frac{n-2}{1-r^2}} = -0.5 \sqrt{\frac{10-2}{1-(-0.5)^2}} \approx -1.633$$

(3) $n = 10$

$\therefore$ there are 8 degrees of freedom

Either For $\alpha = 0.05$, $t_{\frac{\alpha}{2}} = t_{0.025} \approx 2.306$

$\therefore t^*$ does not lie in the critical region.



$$\begin{aligned}
 \text{or } p\text{-value} &= 2P(T > |t^*|) \\
 &\approx 2P(T > 1.633) \\
 &\approx 0.1411 \\
 \therefore p\text{-value} &> \alpha
 \end{aligned}$$

- (4) Using either the critical region test or the p -value, we do not reject H_0 at the 5% level of significance.

We therefore conclude that $\rho = 0$, which means the sets of data are uncorrelated, and X and Y are independent variables.

EXERCISE 1.4

- 1 Perform a hypothesis test on the following bivariate data to determine whether or not the variables are linearly correlated.

<i>Mathematics (x)</i>	13	10	8	14	6	11	10	5	12	13
<i>Physics (y)</i>	21	15	14	20	12	16	9	10	17	12

Consider both 5% and 1% levels of significance, and use both the critical region test and the p -value.

- 2 The cholesterol level in the bloodstream (x) and the resting heart beat (y) of 10 people are:

<i>Cholesterol level (x)</i>	5.32	5.54	5.45	5.06	6.13	5.00	4.90	6.00	6.70	4.75
<i>Resting heart beat (y)</i>	55	48	55	53	74	44	49	68	78	51

Perform a hypothesis test to determine whether or not the variables are linearly correlated. Consider both 5% and 1% levels of significance, and use both the critical region test and the p -value.

- 3 The bivariate data below is sampled from a bivariate normal distribution. X is the number of Sudoku puzzles solved in a three hour period, and Y is the number of logic puzzles solved in a three hour period.

x	5	8	12	15	15	17	20	21	25	27
y	3	11	9	6	15	13	25	15	13	20

Carry out a hypothesis test to determine whether or not the variables are linearly correlated.

Consider both 1% and 5% levels of significance, and use only the p -value.

- 4 Consider the following data sampled from a bivariate normal distribution.

<i>Weight of Mother (x kg)</i>	49	46	48	45	46	42	43	40	52	55	68
<i>Birth weight of child (y kg)</i>	3.8	3.1	2.5	3.0	3.2	2.8	3.1	2.9	3.4	2.9	3.9

Carry out a critical region hypothesis test to determine whether or not the variables X and Y are independent, using:

- a a 5% level of significance b a 1% level of significance.
- 5 A random sample of $n = 16$ paired values from a bivariate normal distribution (X, Y) has correlation coefficient of $r = -0.5$. Does this indicate the variables X and Y are correlated:
- a at the 5% level of significance b at the 1% level of significance?
- 6 A random sample of paired values from a bivariate normal distribution has perfect correlation, so $r = 1$. Explain why this data cannot be used in a hypothesis test for the correlation of X and Y . Do you think a hypothesis test would be needed in this case?

Example 64

A random sample with correlation coefficient $r = 0.5$ is taken from a bivariate normal population (X, Y) . What is the minimum sample size n required to conclude that X and Y are linearly correlated at the 1% level of significance?

We test $H_0: \rho = 0$ against $H_1: \rho \neq 0$ with $r = 0.5$

$$\therefore \text{ the test statistic is } t^* = 0.5 \sqrt{\frac{n-2}{1-(0.5)^2}} \approx 0.577\sqrt{n-2}.$$

For infinitely many degrees of freedom the t -distribution approximates the normal distribution $Z \sim N(0, 1)$

where $t_{\frac{\alpha}{2}} \approx z_{\frac{\alpha}{2}} = z_{0.005} \approx 2.576$.

$$\text{Thus } t^* > t_{\frac{\alpha}{2}} \quad \text{if } 0.577\sqrt{n-2} > 2.576$$

$$\therefore n > 21.9$$

$$\therefore \text{ we require } n \geq 22$$

Try $n = 22$, $\nu = 20$, $t_{0.005} \approx 2.845$, $t^* \approx 2.580$ $\therefore t^* < t_{0.005}$ $\therefore$ not significant.

$n = 23$, $\nu = 21$, $t_{0.005} \approx 2.831$, $t^* \approx 2.644$ $\therefore t^* < t_{0.005}$ $\therefore$ not significant.

$n = 24$, $\nu = 22$, $t_{0.005} \approx 2.818$, $t^* \approx 2.706$ $\therefore t^* < t_{0.005}$ $\therefore$ not significant.

$n = 25$, $\nu = 23$, $t_{0.005} \approx 2.807$, $t^* \approx 2.767$ $\therefore t^* < t_{0.005}$ $\therefore$ not significant.

$n = 26$, $\nu = 24$, $t_{0.005} \approx 2.797$, $t^* \approx 2.827$ $\therefore t^* > t_{0.005}$

$\therefore$ for $n \geq 26$ we would reject $H_0: \rho = 0$ in favour of $H_1: \rho \neq 0$.

$\therefore$ a sample size of at least 26 is required to conclude, at the 1% level of significance, that X and Y are correlated.

- 7** Consider a random sample with correlation coefficient $r = 0.6$ taken from a bivariate normal distribution (X, Y) . What is the minimum sample size n necessary to decide that X and Y are linearly correlated at the 5% level of significance?
- 8** Consider a random sample of 20 data pairs from a bivariate normal distribution (X, Y) . What is the least value of $|r|$ for this sample for which we would conclude, at the 5% level of significance, that X and Y are correlated?

REVIEW SET A

- 1 X_1 , X_2 , and X_3 are random variables, each with mean μ and standard deviation σ .
- Find the mean and standard deviation of:
 - $X_1 + 2X_2 + 3X_3$
 - $2X_1 - 3X_2 + X_3$
 - Find $E([X_1 - X_2]^2)$ given that X_1 and X_2 are independent.
- 2 A student is waiting at a bus stop. He knows that 35% of all the buses going past can take him to school. The other buses go elsewhere.
- Suppose the student will catch the first bus that will take him to school.
 - Find the probability that it will take at most 4 buses for there to be a correct one.
 - Find the *average* number of buses it will take for there to be a correct one.
 - Suppose the student decides that he will catch the *third* bus that would take him to school, because he thinks his friend will be on that bus.
 - Find the probability that he will catch the 7th bus that goes past.
 - Find the average number of buses it will take for the correct bus to arrive.
 - Find the probability that it will take no more than 5 buses for the correct bus to arrive.
- 3 The probability distribution for the random variable X is given in the table alongside. Find the:
- | | | | | | |
|------------|-----|-----|-----|-----|-----|
| $X = x$ | -3 | -1 | 1 | 3 | 5 |
| $P(X = x)$ | c | c | c | c | c |
- value of c
 - mean of X
 - probability that X is greater than the mean
 - variance of X .
- 4 In the Japanese J-League, it is known that 75% of all the footballers prefer to kick with their right leg.
- In a random sample of 20 footballers from the J-League, find the probability that:
 - exactly 14 players prefer to kick with their right leg
 - no more than five prefer to kick with their left leg.
 - In a random sample of 1050 players from the J-League, find the probability that:
 - exactly 70% of players prefer to kick with their right leg
 - no more than 25% prefer to kick with their left leg.
- Hint:** For **b**, use a suitable approximation for the random variable
 X = the number of footballers who prefer to kick with their right leg.
- 5 Suppose $X \sim B(n, p)$ where $x = 0, 1, 2, 3, 4, \dots, n$.
- Show that X has Probability Generating Function $G(t) = (1 - p + pt)^n$.
 - Hence prove that:
 - $E(X) = np$
 - $\text{Var}(X) = np(1 - p)$.
- 6 To estimate the mean number of hours that employees are away from work in a year due to sickness, a sample of 375 people is surveyed. It is found that for last year, the standard deviation for the number of hours lost was 67. Suppose we use this to approximate the standard deviation for this year. Find the probability that the estimate is in error by more than ten hours.



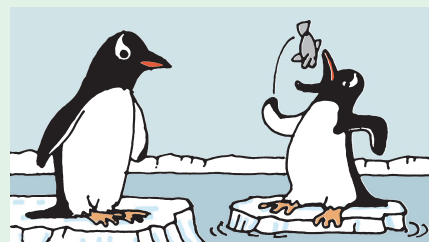
- 7** To work out the credit limit of a prospective credit card holder, a company gives points based on factors such as employment, income, home and car ownership, and general credit history. A statistician working for the company randomly samples 40 applicants and determines the points total for each. His results are:

14 11 13 13 15 12 12 12 10 11 11 11 12 13
 14 13 11 12 14 14 14 13 15 14 11 10 11 16
 11 12 12 10 11 10 10 12 13 13 13 12

- Determine the sample mean $\bar{x}$, and standard deviation s_n .
 - Determine a 95% confidence interval that the company would use to estimate the mean point score for the population of applicants.
- 8** A group of 10 students was given a revision course before their final IB examination. To determine whether the revision course was effective, the students took a test at the beginning and at the end of the course. Their marks are recorded in the table below.

Student	A	B	C	D	E	F	G	H	I	J
Pre-test	12	13	11	14	10	16	14	13	13	12
Post-test	11	14	16	13	12	18	15	14	15	11

- Explain why it would not be appropriate to use the difference between the *means* of these two sets of scores to evaluate whether the revision course was effective.
 - Determine a 90% confidence interval for the mean of the differences in the examination scores. Explain the meaning of your answer.
 - It was hoped that by doing the revision course, the students' scores would improve. Perform an appropriate test at the 5% level of significance to determine whether this was the case.
- 9** Yarni's resting pulse rate was 68 beats per minute for many years. However, using a sensible diet and exercise, she hoped to reduce this rate. After six months, Yarni's mean pulse rate is 65 beats per minute, with a standard deviation of 1.732. These statistics were calculated from 42 measurements. Using the p -value, is there sufficient evidence at a 5% level, to conclude that Yarni's pulse rate has decreased?
- 10** In 2011, the mean weight of a gentoo penguin from a penguin colony was 7.82 kg with standard deviation 1.83 kg. Exactly one year after this data was found, a sample of 48 penguins from the same colony were found to have mean weight 7.55 kg. Is there sufficient evidence, at a 5% level of significance, to suggest that the mean weight in 2012 differs from that in 2011?



- 11**
- Define the population product moment correlation coefficient ρ for the bivariate normal distribution.
 - Deduce that X and Y are independent random variables $\Leftrightarrow \rho = 0$.
 - The bivariate data below compares the height (x cm) and weight (y kg) of 11 men.

Height (x cm)	164	167	173	176	177	178	180	180	181	184	192
Weight (y kg)	68	88	72	96	85	89	71	100	83	97	93

- Calculate the sample product moment correlation coefficient, r .
- At a 5% level of significance, conduct a hypothesis test to determine whether or not the variables X and Y are independent. Use the p -value test only.

12 Suppose distribution X has mean μ and variance σ^2 .

Consider samples $\{x_1, x_2, x_3, \dots, x_n\}$ of fixed size $n > 1$, of random independent values of X .

$$\text{Let } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad S_n^2 = \frac{1}{n} \left\{ \left(\sum_{i=1}^n X_i^2 \right) - n \bar{X}^2 \right\}.$$

- Find $\text{Var}(\bar{X})$.
- Show that $\bar{X}$ is an unbiased estimator of μ .
- Show that S_n^2 is a biased estimator of σ^2 .
- Hence show that $S_{n-1}^2 = \left(\frac{n}{n-1} \right) S_n^2$ is an unbiased estimator of σ^2 .

REVIEW SET B

- 1** At a country fete, there is a game where players pay to randomly choose a card from a table. There is an amount X written on the back of the card. For X being 3 or 6, the player wins \$3 or \$6, but for X being -5 or -1 , the player has to pay an extra \$5 or \$1.

$X = x$	-5	-1	3	6
$P(X = x)$	0.3	0.2	0.2	

- The probability distribution for X is shown in the table.
 - What is the probability of getting a 6 on card X ?
 - What is the expected return per game if the score is the return paid to the player?
 - Explain why organisers should charge \$1 to play this game, rather than 50 cents.
- At the next stall there is a similar game. It uses a different set of cards with outcome Y which has the probability distribution shown.

$Y = y$	-3	2	5
$P(Y = y)$	0.5	0.3	0.2

 - Find the expected return to players for playing this game.
 - Find the expected return for players wishing to play both games simultaneously.
 - How much would you expect the organisers to make if people played games X and Y 500 times each, and the combined game of X and Y 1000 times, and they charged \$1 for every game played?

- 2** A coin is biased so that when it is tossed, the probability of obtaining tails is $\frac{3}{5}$. The coin is tossed 1000 times, and X is the number of tails obtained.

- Find:
 - the mean of X
 - the standard deviation of X .
- Find $P(580 \leq X \leq 615)$ using $X \sim B(1000, \frac{3}{5})$.
- Find $P(579.5 \leq X \leq 615.5)$ using $X \sim N(600, 240)$.
- Explain why the answers in **b** and **c** are so similar.

- 3** By analysing the PDF of the Negative Binomial distribution, prove or disprove: "The Geometric Distribution is a special case of the Negative Binomial distribution."

- 4** Suppose $X \sim \text{Geo}(p)$ where $x = 1, 2, 3, 4, 5, \dots$

- Show that X has PGF given by $G(t) = \frac{pt}{1 - t(1-p)}$ for $|t| < \frac{1}{1-p}$.
- Hence, find $E(X)$ and $\text{Var}(X)$.

- 5** The exponential random variable X has PDF $f(x) = \lambda e^{-\lambda x}$ for all $x \geq 0$. λ is a positive constant.
- Graph $y = f(x)$ for the case $\lambda = 0.8$.
 - For the case $\lambda = 0.8$, find: **i** $E(X)$ **ii** $\text{Var}(X)$
 - Prove that for the general exponential random variable, the median $= \frac{\ln 2}{\lambda}$.
 - Hence show the mean and median for the case $\lambda = 0.8$ on the graph in **a**.
 - Show that the cumulative density function is $F(X) = P(X \leq x) = 1 - e^{-\lambda x}$.
 - Hence find $P(X > 1.3)$ for the case $\lambda = 0.8$.
- 6** A drink manufacturer produces soft drink for sale. Each bottle is advertised as having contents 375 mL. It is known that the machines producing these drinks are set so that the average volume dispensed into each bottle is 376 mL with standard deviation 1.84 mL. The volumes dispensed into each bottle are distributed normally.
- Find the probability that an individual randomly selected bottle has volume less than 373 mL.
 - Find the probability that a randomly selected pack of a dozen bottles has an average volume less than the advertised amount.
 - Government regulations are set to ensure that companies meet their advertising claims. This company can choose to be tested under either of the following rules:
 - A randomly selected bottle must contain no less than 373 mL. *or*
 - A randomly selected pack of 12 bottles must have average contents no less than the advertised amount.

Explain clearly by which method the company would prefer to be tested by the Government authority.
 - Suppose the company chose to be tested using method **II** above. The company wants less than 0.1% chance of being fined by the Government Authority for failing to meet the requirement.
- Find, to the nearest mL, what the setting should be for the average volume dispensed into each bottle, assuming the standard deviation is unchanged.
- 7** The manufacturer of the breakfast cereal Maxiweet knows that the net contents of each packet has variance 151.4 grams². A sample of 120 packets is chosen at random, and their mean weight is found to be 596.7 grams/packet.
- Construct a 95% confidence interval for the true mean of the population. Interpret your answer.
 - The manufacturer claims that the net weight of each packet is 600 g. Is there sufficient evidence to support the manufacturer's claim at the 5% level of significance?
 - Find the sample size required if the manufacturer wishes to be 95% confident that the sample mean differs from the population mean by less than 2 grams.
- 8** The house prices in a large town are normally distributed. A real estate agent claims that the mean house price is €438 000. To test the real estate agent's claim, 30 recently sold houses were randomly selected and the mean price $\bar{x}$ was calculated. The standard deviation for this sample was €23 500. What values of $\bar{x}$, to the nearest €500, would support the agent's claim at a 2% level of significance?

- 9 The covariance of random variables X and Y is defined as:

$$\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

- a Prove that $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
 - b Hence, prove that if X and Y are independent, then $E(XY) = E(X)E(Y)$
 - c Given that X and Y are independent, expand and simplify $\text{Cov}(X, X - Y)$.
- 10 A machine is used to fill packets with rice. The contents of each packet weighs X grams. X is normally distributed with mean μ grams and standard deviation 3.71 grams. The mean weight μ is stated to be 500 g. To check this statement, a sample of 13 packets is selected, and the mean contents $\bar{x}$ is calculated. The hypotheses for the test are $H_0: \mu = 500$ g and $H_1: \mu \neq 500$ g. The critical region is defined by $\{\bar{x} < 498\} \cup \{\bar{x} > 502\}$.
- a What is the nature of the distribution of $\bar{X}$?
 - b What is the meaning of the critical region?
 - c Show that the significance level for the test is approximately 0.0519.
 - d Find the probability of a Type II error with this test given that the true value of μ is 498.4 grams.
- 11 In 2011 a market gardener found that the weight of his tomatoes was normally distributed with mean 106.3 g and standard deviation 12.41 g. In 2012 he used a different fertiliser and found that a randomly selected sample of 65 tomatoes had mean weight 110.1 g. Assuming the population standard deviation in 2012 is again 12.41 g, is there sufficient evidence at a 1% level, to suggest that the mean weight of a tomato has increased?
- 12 A population X has mean μ and variance σ^2 . Random samples $\{x_1, x_2\}$ of size 2 of random independent values are taken from X .
- a Show that $T_1 = \frac{3X_1 + 7X_2}{10}$ is an unbiased estimator of μ .
 - b Show that $T_2 = \frac{3X_1 + 2X_2}{5}$ is also an unbiased estimator of μ .
 - c A random sample $\{2.6, 5.3\}$ is taken from X . Lucy uses T_1 and Beth uses T_2 to estimate μ . Eve says that estimator T_1 is more efficient than estimator T_2 .
 - i Calculate Lucy's estimate.
 - ii Calculate Beth's estimate.
 - iii Is Eve correct? Give reasons for your answer.
 - d Let $T_3 = \frac{aX_1 + bX_2}{c}$ be an estimator of μ for constants a, b , and $c \in \mathbb{R}^+$. State the required conditions on a, b , and c for T_3 to be an unbiased estimator of μ .



REVIEW SET C

- 1** X_1 , X_2 , and X_3 are independent random variables where $X_1 \sim N(2, \frac{1}{8})$, $X_2 \sim N(3, \frac{1}{16})$, and $X_3 \sim N(a, b)$.
- If $Y = 2X_3 - 2X_2 - X_1$, find:
 - $E(Y)$
 - $\text{Var}(Y)$
 - Given that $E(Y) = 0$ and $\text{Var}(Y) = 1$, find the values of a and b .
 - State the nature of the distribution of Y .
 - Find $P(X_3 \geq 8b)$.
- 2** Pierre runs a game at a fair, where each player is guaranteed to win €10. Players pay a certain amount each time they roll an unbiased die, and *must* keep rolling until a '6' occurs. When a '6' occurs, Pierre gives the player €10 and the game concludes. On average, Pierre wishes to make a profit of €2 per game. How much does he need to charge for each roll of the die?
- 3** It is known that the probability of a journalist making no errors on each page is q .
- State the distribution of the random variable X that defines the number of errors made per page by that journalist.
 - Find the probability, in terms of q , that the journalist makes per page:
 - one error
 - more than one error.
 - The journalist gets a bonus of \$10 for a page with no errors or \$1 for just one error on the page, but loses \$8 from their pay if there is more than one error on a page.
 - Draw a probability distribution table for the random variable Y which describes the returns for the journalist for making different numbers of errors on a page.
 - Find $E(Y)$ in terms of q .
 - Find the smallest value of q , $0 \leq q \leq 1$, such that the journalist will receive an overall bonus. Give your answer to three decimal places.
- 4** The weekly demand for petrol (in thousands of kilolitres) at a service station is a continuous random variable with probability density function $f(x) = ax^3 + bx^2$, $0 \leq x \leq 1$.
- If the mean weekly demand is 700 kilolitres, determine the values of a and b .
 - Suppose the service station has a storage capacity of 950 kilolitres. Find the probability that the service station will run out of petrol in any given week.
- 5** The PGF for the Negative Binomial variable X is $G(t) = \left(\frac{1-p}{1-pt}\right)^r$ for $|t| < \frac{1}{p}$ and $r = 0, 1, 2, 3, 4, \dots$
- The mean of X is $\mu = \frac{pr}{1-p}$.
- Show that $p = \frac{\mu}{\mu + r}$.
 - Hence, show that $G(t) = \frac{1}{\left(1 + \frac{\mu(1-t)}{r}\right)^r}$.
 - Find $\lim_{r \rightarrow \infty} G(t)$ and interpret your result.
 - Copy and complete: $\text{Po}(m) = \lim_{r \rightarrow \infty} \text{NB}(r, \dots)$.

- 6** The weight W of gourmet sausages produced by Hans is normally distributed with mean 63.4 g and variance 40.1 g^2 .

Suppose $\bar{W}$ is the mean weight of 10 randomly selected sausages.

- Find the probability that a randomly chosen sausage will have a weight of 60 grams or less.
- State the distribution of $\bar{W}$.
- Hence, calculate the probability that $\bar{W} \leq 60 \text{ g}$. Interpret your result.
- How large should a sample size be so there is less than 1% chance that the mean weight of sausages in the sample will be less than 60 g?



- 7** Romano's Constructions are concerned about the number of its employees having headaches at the end of their daily work period. They suspect that the headaches may be due to high systolic blood pressure. High systolic blood pressure is generally diagnosed when a blood pressure test is more than 140 mm Hg.

A doctor measures the systolic blood pressure of a random sample of the company employees, and finds that the sample standard deviation is 11.2 mm Hg.

The doctor calculates the 95% confidence interval for the population mean to be $139.91 < \mu < 147.49$.

- Write down, in terms of the sample mean $\bar{x}$, a general expression for the 95% CI for μ .
 - Hence find $\bar{x}$.
 - Explain why the CI in **a** cannot be used to find the sample size n without resorting to trial and error.
 - Use trial and error to find the sample size, given that $n > 32$.
- 8** A school claims to be able to teach anglers how to fish better and catch more fish. In order to test this hypothesis, the school recorded the number of fish caught by a random sample of nine anglers at a local jetty in a three hour period. The school then gave the anglers a free course to help them with their fishing. After the fishing course was completed, they again recorded the number of fish caught by the anglers at the same jetty in the same time period, at the same time of day. The results were:

Angler	A	B	C	D	E	F	G	H	I
Number of fish caught before	24	23	22	30	41	30	33	18	15
Number of fish caught after	36	32	40	27	32	34	33	28	19

- Test at the 5% level, whether the fishing school's claim is indeed correct. State the type of error you could make.
 - Find the 90% confidence interval for the mean difference of the two sets of scores, and interpret the meaning of your answer.
- 9** Last year a ten-pin bowler consistently bowled a score of around 200 each game, with a standard deviation of 11.36.

Her last 35 scores have had mean 196.4, and she is convinced that her form is below that of last year.

Assuming her scores are normally distributed, test her claim at a 2% level, using the critical region.

- 10** A random variable X has a Poisson distribution with unknown mean m . We wish to test the hypothesis $H_0: m = 3$ against $H_1: m \neq 3$.
To do this, a random sample of 15 independent values $\{x_1, x_2, x_3, \dots, x_{15}\}$ is taken from X , with replacement.

The decision rule is: accept H_0 if $34 \leq \sum_{i=1}^{15} x_i \leq 56$, otherwise reject H_0 .

- Define the critical region of $S = \sum_{i=1}^{15} x_i$.
 - Find the Type I error and calculate its probability.
 - If the true value of m is 3.4, what is the probability of a Type II error?
- 11** A random sample of 27 data pairs is taken from a bivariate normal distribution (X, Y) . What is the greatest value of $|r|$ for this sample for which we can conclude that, at a 5% level of significance, X and Y are independent? Give your answer correct to 4 significant figures.

- 12** Suppose independent random samples of independent values are taken from the same population with unknown mean μ and unknown variance σ^2 .

For a sample of size n , the sample variance is defined by $s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

Sample A of size 8 has sample variance $s_A^2 = 4.2$.

Sample B of size 22 has sample variance $s_B^2 = 5.1$.

- Use each of the samples A and B to find two unbiased estimates of σ^2 .

- Let $t_1 = \frac{8s_A^2 + 22s_B^2}{30}$.

- Calculate the estimate t_1 of σ^2 .

- Is t_1 a biased or unbiased estimate of σ^2 ? Explain your answer.

- Let $t_2 = \frac{as_A^2 + bs_B^2}{c}$.

State the required conditions on $a, b, c \in \mathbb{R}^+$ for t_2 to be an unbiased estimate of σ^2 .

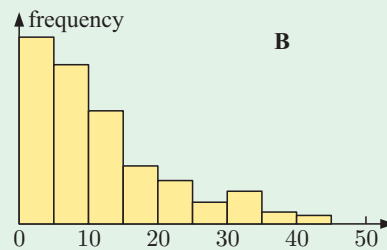
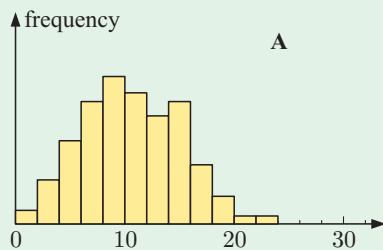
REVIEW SET D

- 1** A soft drink manufacturer produces small and large bottles of drink. The volumes of both sizes of drink are normally distributed with means and standard deviations given in the table alongside.

	μ (mL)	σ (mL)
small bottles	338	3
large bottles	1010	12

- Find the probability that one large bottle selected at random will contain more than the combined contents of three smaller bottles selected at random.
 - Find the probability that one large bottle selected at random will contain three times more than one smaller bottle selected at random.
- 2** Patients arrive at random to the hospital Emergency room, at a rate of 14 per hour. Find the probability that:
- exactly five patients arrive between 9:00 am and 9:45 am
 - fewer than seven patients will arrive between 10:00 am and 10:30 am.

- 3** The random variable $X \sim B(n, p)$ has $E(X) = 8$ and $\text{Var}(X) = 6$.
- Find n and p .
 - Find:
 - $P(X \geq 4)$
 - $P(7.9 \leq \bar{X} \leq 8.1)$
- 4** The normal PDF with parameters μ and σ^2 is $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $x \in]-\infty, \infty[$.
Use $f(x)$ to find the *exact* value of $\int_{-\infty}^{\infty} e^{-\frac{1}{8}(x-9)^2} dx$.
- 5** The PGF for the continuous random variable X with a Gamma distribution is $G(t) = (1 - \beta \ln t)^{-\alpha-1}$, where α and β are parameters.
- Assuming $\mu = G'(1)$ and $\text{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$, show that the mean of the Gamma distribution is $\mu = (\alpha + 1)\beta$.
 - Hence determine the variance of the Gamma distribution.
- 6** The random variable X has a normal distribution with mean μ . A randomly selected sample of size 15 is taken, and it is found that $\sum_{i=1}^{15} (x_i - \bar{x})^2 = 230$.
- Find the sample variance for this sample.
 - Find an unbiased estimate of the population variance for the random variable X .
 - A confidence interval for μ (*not* the 95% confidence interval) taken from this sample is $]124.94, 129.05[$.
 - Find the 95% confidence interval for μ taken from this sample.
 - Determine the confidence level for the confidence interval $]124.94, 129.05[$.
- 7** In order to estimate the copper content of a potential mine, drill core samples are used. All of the drill core is crushed and well mixed before samples are removed.
Suppose X is the copper content in grams per kilogram of core, and that X is normally distributed with mean $\mu = 11.4$ and $\sigma = 9.21$.
Samples of size 9 are randomly chosen from the X -distribution.
Histograms **A** and **B** (shown below) are for the X and $\bar{X}$ distributions, not necessarily in that order.



- Which of these histograms is the histogram for X ? Explain your answer.
- For the $\bar{X}$ -distribution, find the mean $\mu_{\bar{X}}$ and variance $\sigma_{\bar{X}}^2$.
- Find $P(\bar{X} \geq 12)$.
- Another drill core is obtained and $\bar{Y}$ is the average copper content of n random samples. With the assumption that $\sigma_X = \sigma_Y = 9.21$, find how large n should be for $\bar{Y}$ to be within ± 3 grams per kg of $\mu_{\bar{Y}}$ with 95% probability.

- 8** A large business uses hundreds of light bulbs each year. They currently use Brand A bulbs which have a mean life of 546 hours. A supplier of another brand, Brand B, will supply bulbs at the same price as Brand A, and claims that these bulbs have a mean life in excess of 546 hours.



The business will purchase Brand B bulbs if, when they test a random sample of 50 bulbs, the supplier's claim is supported at a 5% level of significance. When the 50 bulbs were tested, the mean life was 563 hours with a variance of 3417 hours². Is the supplier's claim acceptable?

- 9** The random variable X has a geometric distribution with unknown parameter p . Theo wishes to test the hypothesis $H_0: p = 0.25$ against $H_1: p \neq 0.25$.

A random sample $\{x_1, x_2, x_3, \dots, x_{12}\}$ of independent values is taken from X , with replacement.

The critical region is defined as $\left\{ \sum_{i=1}^{12} x_i \leq 31 \right\} \cup \left\{ \sum_{i=1}^{12} x_i \geq 75 \right\}$.

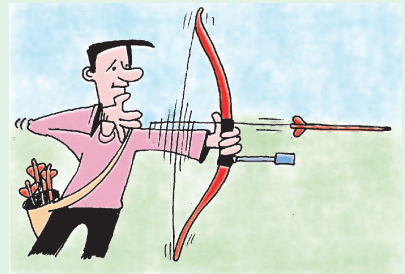
- If $S = \sum_{i=1}^{12} x_i$, what is the distribution of S under H_0 ?
 - What is the acceptance region for S under the null hypothesis H_0 ?
 - Calculate the level of significance α for this test.
 - The true value of p is $p = 0.2$. Calculate the power of the test.
- 10** Quickchick grow chickens to sell to a supermarket chain. However, the buyers believe that the chickens are supplied underweight. As a consequence they consider the hypotheses:
- H_0 : Quickchick is not supplying underweight chickens
 H_1 : Quickchick is supplying underweight chickens.

What conclusion would result in: **a** a type I error **b** a type II error?

- 11** An archer shoots 10 arrows at a target from each of 12 different positions. The table below shows the distance of each position from the target, and how many shots were successful.

Position	A	B	C	D	E	F	G	H	I	J	K	L
Distance from target (x m)	20	25	15	35	40	55	30	45	60	80	65	70
Hits (y)	9	8	8	8	7	6	9	7	4	2	3	3

- Draw a scatter plot of the data.
- Calculate the sample correlation coefficient r .
- Comment on the strength of the correlation between X and Y .
- Is there a causal relationship between the variables?
- Find the equation of the regression line of Y on X .
- Predict the number of hits made from 50 m if the archer shoots 100 arrows.
- Should the regression line be used to predict the number of hits when arrows are fired from 100 m? Explain your answer.
- Find the equation of the regression line of X on Y .



- 12** X is a random variable with unknown mean μ and unknown variance σ^2 .

Consider samples of size 2, $\{x_1, x_2\}$, of independent values taken from X .

Let $T_1 = \frac{2X_1 + X_2}{3}$ and $T_2 = aX_1 + (1 - a)X_2$ for some constant $a \in [0, 1]$.

- a** For each estimator T_1 and T_2 , determine whether it is a biased or unbiased estimator of μ .
- b** Calculate $\text{Var}(T_1)$ and $\text{Var}(T_2)$.
- c** For which value(s) of a is T_2 a more efficient estimator than T_1 ?
- d** For which value of $a \in [0, 1]$ is the estimator T_2 most efficient?

- 13** Suppose X and Y are independent random variables with $E(X) = \mu_X$, $\text{Var}(X) = \sigma_X^2$, $E(Y) = \mu_Y$, and $\text{Var}(Y) = \sigma_Y^2$.

A random sample of size n is taken from X , and the sample mean $\bar{x}$ and $s_X^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$ are calculated.

Similarly, a sample of size m is taken from Y , and the sample mean $\bar{y}$ and $s_Y^2 = \frac{\sum_{i=1}^m (y_i - \bar{y})^2}{m - 1}$ are calculated.

Let $U = 3X - 5Y$.

- a** Find $E(U)$ and $\text{Var}(U)$ in terms of μ_X , μ_Y , σ_X^2 , and σ_Y^2 where appropriate.
- b** If $aS_X^2 + bS_Y^2$ is an unbiased estimate of $\text{Var}(U)$, write down the values of the constants a and b .

THEORY OF KNOWLEDGE

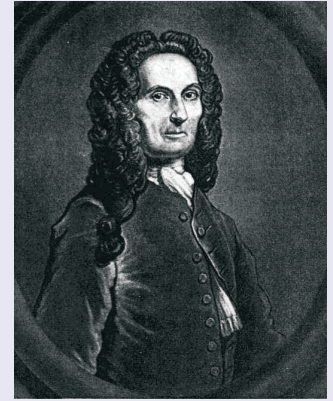
THE CENTRAL LIMIT THEOREM

The French mathematician **Abraham de Moivre** is most famous for his formula linking complex numbers and trigonometry, but he was also responsible for the early development of the Central Limit Theorem. In an article from 1733, he used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin.

The work of de Moivre was extended in 1812 by country-man **Pierre-Simon Laplace**, but the theorem was not formalised and rigorously proven until the early 20th century work of the Russian mathematicians **Pafnuty Chebyshev**, **Andrey Markov**, and **Aleksandr Lyapunov**.

In 1889, **Sir Francis Galton** wrote about the normal distribution, with particular relevance to the Central Limit Theorem:

“I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the law of frequency of error. The law would have been personified by the Greeks if they had known of it. It reigns with serenity and complete self-effacement amidst the wildest confusion. The larger the mob, the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.”



Abraham de Moivre

- 1 How is mathematical order linked to our sense of beauty?

The name “Central Limit Theorem” was first used in 1920 by the Hungarian mathematician **George Pólya** (1887-1985). He used the word “central” because of the importance of the theorem in probability theory.

- 2 What makes a theorem “fundamental”? Could the Central Limit Theorem reasonably be referred to as the Fundamental Theorem of Statistics?
- 3 Discuss the statement: “Without the Central Limit Theorem, there could be no statistics of any value within the human sciences.”

The philosophies of **rationalism** and **empiricism** concern the extent to which we are dependent upon sense experience in order to gain knowledge.

Rationalists claim that knowledge can be gained through logic independent of sense experience. They argue about the limitations of what sense experience can provide, and how reason in other forms contributes additional information about the world.

By contrast, empiricists claim that sense experience is most important for knowledge, arguing that we cannot regard something as knowledge if experience cannot provide it.

- 4 Is it more important to rationalise the Central Limit Theorem by mathematical proof, or confirm its truth by empirical application?

THEORY OF KNOWLEDGE

POPULATION PARAMETERS

In previous courses we have seen that data collected from a population may be **qualitative** or **quantitative**.

However, there are other properties of parameters that can affect how we need treat them. We may need to sample for data in a particular way, or word questions carefully so the person responding understands the context in which the question is being asked.

Data is **objective** if the result of its measurement is clear.

For example:

- The number of cousins you have is objective, quantitative, and discrete.
- A person's height, and weight are all objective, quantitative, and continuous.
- Either a person is infected with HIV, or they are not, so this parameter is objective and qualitative.

By contrast, data is **subjective** if the result of its measurement depends on individual interpretation, and is relative to the individual's own experiences.

For example:

- The colour of a person's shirt is subjective and qualitative. For some colours like black and white, people will agree on the result, but people may argue about whether a shirt is red or pink, or perhaps blue or purple.
- A person's mood is subjective and qualitative. We can ask a person to rate themselves on a scale of sad to happy, or confident to fearful, but some states such as angry do not necessarily have defined opposites. Moods are also **transitive** in the sense that they can change rapidly, so a data set correct at the time of measurement may be very different to that which would be measured from the same sample of people soon after.



- 1 Is subjective data just the product of a vague question? For example, can the question "How many friends do you have?" be refined so it is no longer subjective?
- 2 Are statistics gathered on transitive parameters less meaningful?
- 3 What do you regard as the most important things in life? Are these things qualitative or quantitative? Are they objective or subjective? Are they transitive?
- 4 To what extent is "well-being" a social construct?
- 5 What sort of social consciousness is necessary for a subjective response to fit into the broader context of society?
- 6 Does the ability to test only certain parameters in a population affect the way knowledge claims in the human sciences are valued?

The measurement of performance poses particular challenges for statisticians because a contextual framework needs to be set in which the performance is measured.

For example, the performance of an athlete may be measured by how fast they run, how high they jump, or how far they throw, but in doing this we require a standard for comparison. Do we compare the results of an athlete to those of other athletes of the same age or the same gender or the same race

or the same body physique, or against all athletes in the world? Or is it more important to compare the athlete's results against themselves and what is physically achievable for the individual?

- 7** How can we measure evolutionary success? You may wish to think about number of progeny, life-span of a species, and diversity within a species.
- 8** How can we measure the performance of an organisation? Should the performance be measured by its owners, its managers, its employees, or its customers? How does profit balance against service, or company mission?

Worked Solutions

EXERCISE A

1 $X \sim N(5, 22)$. $\sigma^2 = 22$
 $\therefore \sigma = \sqrt{22}$

a $P(X = 24) = P(23.5 \leq X < 24.5)$
 $\approx 0.000\,023\,9$

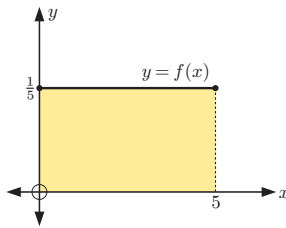
b $P(X = 30) = P(29.5 \leq X < 30.5)$
 $\approx 0.000\,000\,060\,8$

c $P(X = 11) = P(10.5 \leq X < 11.5)$
 ≈ 0.0376

2 a $E(X) = -1(\frac{1}{2}) + 1(\frac{1}{3}) + 3(\frac{1}{6})$
 $= -\frac{1}{2} + \frac{1}{3} + \frac{1}{2}$
 $= \frac{1}{3}$

b $X \sim U(0, 5)$

$$\begin{aligned} E(X) &= \int_0^5 x f(x) dx \\ &= \int_0^5 x(\frac{1}{5}) dx \\ &= \frac{1}{5} \left[\frac{x^2}{2} \right]_0^5 \\ &= \frac{1}{5} \times \frac{25}{2} \\ &= 2.5 \end{aligned}$$



x	-3	-1	2	3
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

a $E(X) = \sum x_i p_i$
 $= -3(\frac{1}{6}) - 1(\frac{1}{3}) + 2(\frac{1}{3}) + 3(\frac{1}{6})$
 $= -\frac{1}{2} - \frac{1}{3} + \frac{2}{3} + \frac{1}{2}$
 $= \frac{1}{3}$

Possible X^2 values are 1, 4, and 9.

$P(X^2 = 1) = P(X = -1 \text{ or } 1)$
 $= P(X = -1)$
 $= \frac{1}{3}$

$P(X^2 = 4) = P(X = 2 \text{ or } -2)$
 $= P(X = 2)$
 $= \frac{1}{3}$

$P(X^2 = 9) = P(X = 3 \text{ or } -3)$
 $= \frac{1}{3}$

$\therefore E(X^2) = \sum x_i^2 p_i$
 $= 1(\frac{1}{3}) + 4(\frac{1}{3}) + 9(\frac{1}{3})$
 $= \frac{14}{3}$

b $E(X^2) = \frac{14}{3}$ and $(E(X))^2 = \frac{1}{9}$
 $\therefore E(X^2) \neq (E(X))^2$

4 a i

x	0	1	2
Probability	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
x^2	0	1	4
Probability	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

For example: $P(X^2 = 4) = P(X = 2 \text{ or } -2)$
 $= \frac{1}{4} + 0 = \frac{1}{4}$

ii

y	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
y^2	1	4	9	16	25	32
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

iii

$x + y$	1	2	3	4	5	6	7	8
Probability	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{24}$

For example: $P(X + Y = 5)$
 $= P(X = 0, Y = 5 \text{ or } X = 1, Y = 4,$
 $\text{or } X = 2, Y = 3)$
 $= (\frac{1}{4})(\frac{1}{6}) + (\frac{1}{2})(\frac{1}{6}) + (\frac{1}{4})(\frac{1}{6})$
 $= \frac{1}{6}$

$(x + y)^2$	1	4	9	16	25	36	49	64
Probability	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{24}$

iv

$4x - 2y$	-12	-10	-8	-6	-4	-2	0	2	4	6
Probability	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{24}$

$(4x - 2y)^2$	0	4	16	36	64	100	144
Probability	$\frac{1}{8}$	$\frac{7}{24}$	$\frac{5}{24}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{24}$

For example: $P((4X - 2Y)^2 = 4)$
 $= P(4X - 2Y = 2 \text{ or } -2)$
 $= \frac{1}{8} + \frac{1}{6}$
 $= \frac{7}{24}$

v

xy	0	1	2	3	4	5	6	8	10	12
Prob.	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$

$(xy)^2$	0	1	4	9	16	25	36	64	100	144
Prob.	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$

b i $E(X) = 0(\frac{1}{4}) + 1(\frac{1}{2}) + 2(\frac{1}{4}) = 1$

$\text{Var}(X) = E(X^2) - [E(X)]^2$
 $= 0(\frac{1}{4}) + 1(\frac{1}{2}) + 4(\frac{1}{4}) - 1^2$
 $= \frac{1}{2}$

ii $E(Y) = 1(\frac{1}{6}) + 2(\frac{1}{6}) + 3(\frac{1}{6}) + 4(\frac{1}{6}) + 5(\frac{1}{6}) + 6(\frac{1}{6})$
 $= 3.5$

$\text{Var}(Y) = E(Y^2) - [E(Y)]^2$
 $= 1(\frac{1}{6}) + 4(\frac{1}{6}) + 9(\frac{1}{6}) + 16(\frac{1}{6}) + 25(\frac{1}{6})$
 $+ 36(\frac{1}{6}) - 3.5^2$
 ≈ 2.92

$$\begin{aligned} \text{iii } E(X + Y) &= 1\left(\frac{1}{24}\right) + 2\left(\frac{1}{8}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) \\ &\quad + 6\left(\frac{1}{6}\right) + 7\left(\frac{1}{8}\right) + 8\left(\frac{1}{24}\right) \\ &= 4.5 \end{aligned}$$

$$\begin{aligned} \text{Var}(X + Y) &= E((X + Y)^2) - [E(X + Y)]^2 \\ &= 1\left(\frac{1}{24}\right) + 4\left(\frac{1}{8}\right) + 9\left(\frac{1}{6}\right) + 16\left(\frac{1}{6}\right) + 25\left(\frac{1}{6}\right) \\ &\quad + 36\left(\frac{1}{6}\right) + 49\left(\frac{1}{8}\right) + 64\left(\frac{1}{24}\right) - 4.5^2 \\ &\approx 3.42 \end{aligned}$$

$$\begin{aligned} \text{iv } E(4X - 2Y) &= -12\left(\frac{1}{24}\right) - 10\left(\frac{1}{24}\right) - 8\left(\frac{1}{8}\right) - 6\left(\frac{1}{8}\right) - 4\left(\frac{1}{6}\right) \\ &\quad - 2\left(\frac{1}{6}\right) + 0\left(\frac{1}{8}\right) + 2\left(\frac{1}{8}\right) + 4\left(\frac{1}{24}\right) + 6\left(\frac{1}{24}\right) \\ &= -3 \end{aligned}$$

$$\begin{aligned} \text{Var}(4X - 2Y) &= E((4X - 2Y)^2) - [E(4X - 2Y)]^2 \\ &= 0\left(\frac{1}{8}\right) + 4\left(\frac{7}{24}\right) + 16\left(\frac{5}{24}\right) + 36\left(\frac{1}{6}\right) + 64\left(\frac{1}{8}\right) \\ &\quad + 100\left(\frac{1}{24}\right) + 144\left(\frac{1}{24}\right) - (-3)^2 \\ &\approx 19.7 \end{aligned}$$

$$\begin{aligned} \text{v } E(XY) &= 0\left(\frac{1}{4}\right) + 1\left(\frac{1}{12}\right) + 2\left(\frac{1}{8}\right) + 3\left(\frac{1}{12}\right) + 4\left(\frac{1}{8}\right) + 5\left(\frac{1}{12}\right) \\ &\quad + 6\left(\frac{1}{8}\right) + 8\left(\frac{1}{24}\right) + 10\left(\frac{1}{24}\right) + 12\left(\frac{1}{24}\right) \\ &= 3.5 \\ \text{Var}(XY) &= E((XY)^2) - [E(XY)]^2 \\ &= 0\left(\frac{1}{4}\right) + 1\left(\frac{1}{12}\right) + 4\left(\frac{1}{8}\right) + 9\left(\frac{1}{12}\right) + 16\left(\frac{1}{8}\right) + 25\left(\frac{1}{12}\right) \\ &\quad + 36\left(\frac{1}{8}\right) + 64\left(\frac{1}{24}\right) + 100\left(\frac{1}{24}\right) + 144\left(\frac{1}{24}\right) - 3.5^2 \\ &= 10.5 \end{aligned}$$

$$\begin{aligned} \text{c i } E(X + Y) &= 4.5 \\ E(X) + E(Y) &= 1 + 3.5 = 4.5 \quad \checkmark \\ \text{Var}(X + Y) &\approx 3.42 \\ \text{Var}(X) + \text{Var}(Y) &\approx \frac{1}{2} + 2.917 \approx 3.42 \quad \checkmark \end{aligned}$$

$$\begin{aligned} \text{ii } E(4X - 2Y) &= -3 \\ 4E(X) - 2E(Y) &= 4 \times 1 - 2 \times 3.5 \\ &= 4 - 7 \\ &= -3 \quad \checkmark \\ \text{Var}(4X - 2Y) &\approx 19.7 \\ 16\text{Var}(X) + 4\text{Var}(Y) &\approx 16\left(\frac{1}{2}\right) + 4(2.917) \\ &\approx 19.7 \quad \checkmark \end{aligned}$$

$$\begin{aligned} \text{iii } E(XY) &= 3.5 \\ E(X)E(Y) &= 1 \times 3.5 = 3.5 \quad \checkmark \end{aligned}$$

$$\text{5 } E(X) = 3, E(Y) = 2, \text{Var}(X) = \frac{3}{2}, \text{Var}(Y) = \frac{5}{4}$$

If X and Y are independent,

$$\begin{array}{ll} \text{a } E(X + Y) & \text{b } E(XY) \\ = E(X) + E(Y) & = E(X)E(Y) \\ = 3 + 2 & = 3 \times 2 \\ = 5 & = 6 \end{array}$$

$$\begin{aligned} \text{c } \text{Var}(2X - 3Y + 6) &= 4\text{Var}(X) + 9\text{Var}(Y) + \text{Var}(6) \\ &= 4\left(\frac{3}{2}\right) + 9\left(\frac{5}{4}\right) + 0 \\ &= 17.25 \end{aligned}$$

$$\begin{aligned} \text{d } E(5XY) &= 5E(XY) \\ &= 5 \times 6 \\ &= 30 \end{aligned}$$

If X and Y are dependent, $E(X + Y) = 5$.

However, $E(XY)$, $\text{Var}(2X - 3Y + 6)$, and $E(5XY)$ can only be determined if X and Y are independent.

$$\text{6 a } \begin{array}{|c|c|c|c|c|} \hline x & 0 & 1 & 2 & 3 \\ \hline \text{Probability} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \hline \end{array}$$

$$\begin{aligned} E(X) &= 0\left(\frac{1}{4}\right) + 1\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{4}\right) \\ \therefore E(X) &= \frac{3}{2} \end{aligned}$$

$$\begin{array}{|c|c|c|c|c|} \hline y = x^2 & 0 & 1 & 4 & 9 \\ \hline \text{Probability} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \hline \end{array}$$

$$\begin{aligned} \therefore E(X^2) &= 0\left(\frac{1}{4}\right) + 1\left(\frac{1}{4}\right) + 4\left(\frac{1}{4}\right) + 9\left(\frac{1}{4}\right) \\ \therefore E(Y) &= \frac{7}{2} \end{aligned}$$

$$\begin{array}{|c|c|c|c|c|} \hline xy = x^3 & 0 & 1 & 8 & 27 \\ \hline \text{Probability} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \hline \end{array}$$

$$\begin{aligned} \therefore E(XY) &= 0\left(\frac{1}{4}\right) + 1\left(\frac{1}{4}\right) + 8\left(\frac{1}{4}\right) + 27\left(\frac{1}{4}\right) \\ &= 9 \\ E(X)E(Y) &= \frac{3}{2} \times \frac{7}{2} = \frac{21}{4} \neq E(XY) \end{aligned}$$

$$\text{b } \begin{array}{|c|c|c|c|c|} \hline x & -1 & 0 & 1 & 2 \\ \hline \text{Probability} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \hline \end{array}$$

$$\begin{aligned} E(X) &= -1\left(\frac{1}{4}\right) + 1\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) \\ \therefore E(X) &= \frac{1}{2} \end{aligned}$$

$$\begin{array}{|c|c|c|c|} \hline y = x^2 & 0 & 1 & 4 \\ \hline \text{Probability} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \hline \end{array}$$

$$\begin{aligned} E(Y) &= 1\left(\frac{1}{2}\right) + 4\left(\frac{1}{4}\right) \\ \therefore E(Y) &= 1\frac{1}{2} \end{aligned}$$

$$\begin{array}{|c|c|c|c|c|} \hline xy = x^3 & -1 & 0 & 1 & 8 \\ \hline \text{Probability} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \hline \end{array}$$

$$\begin{aligned} \therefore E(XY) &= -1\left(\frac{1}{4}\right) + 1\left(\frac{1}{4}\right) + 8\left(\frac{1}{4}\right) \\ &= 2 \\ \therefore E(X)E(Y) &= \frac{3}{4} \neq E(XY) \end{aligned}$$

$$\begin{aligned} \text{7 a } E(X) &= 3.8, E(Y) = 5.7 \\ E(3X - 2Y) &= 3E(X) - 2E(Y) \\ &= 3 \times 3.8 - 2 \times 5.7 \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Var}(3X - 2Y) &= 9\text{Var}(X) + 4\text{Var}(Y) \\ &= 9 \times 0.323^2 + 4 \times 1.02^2 \\ &= 5.100561 \end{aligned}$$

$\therefore$ the standard deviation of $3X - 2Y$ is $\sqrt{5.100561}$ which is ≈ 2.26 .

b By **Theorem 8**, the linear combination $3X - 2Y$ of independent normally distributed random variables X and Y is also normally distributed
 $\therefore 3X - 2Y \sim N(0, 2.25844^2)$
 Thus $P(3X - 2Y > 3)$
 ≈ 0.0920 {using technology}

8 $X \sim N(\mu, \sigma^2)$

Now $P(X \geq 80) = 0.1$ and $P(X \geq 65) = 0.3$

$\therefore P(X < 80) = 0.9$ and $P(X < 65) = 0.7$

$\Rightarrow P\left(\frac{X - \mu}{\sigma} < \frac{80 - \mu}{\sigma}\right) = 0.9$ and

$P\left(\frac{X - \mu}{\sigma} < \frac{65 - \mu}{\sigma}\right) = 0.7$

$\Rightarrow P\left(Z < \frac{80 - \mu}{\sigma}\right) = 0.9$ and $P\left(Z < \frac{65 - \mu}{\sigma}\right) = 0.7$

$\Rightarrow \frac{80 - \mu}{\sigma} \approx 1.28155$ and $\frac{65 - \mu}{\sigma} \approx 0.52440$

$\Rightarrow 80 - \mu \approx 1.28155\sigma$ and $65 - \mu \approx 0.52440\sigma$

Solving simultaneously $\mu \approx 54.6$ and $\sigma \approx 19.8$

- 9** Let A be the random variable for the weight of an adult and C be the random variable for the weight of a child.

So, $A \sim N(81, 11^2)$ and $C \sim N(48, 4^2)$.

Consider the random variable S where

$$S = A_1 + A_2 + A_3 + A_4 + C_1 + C_2 + C_3$$

Now $E(S) = E(A_1) + E(A_2) + E(A_3) + \dots + E(C_3)$

$$= 4 \times 81 + 3 \times 48$$

$$= 468$$

and $\text{Var}(S) = \text{Var}(A_1) + \text{Var}(A_2) + \dots + \text{Var}(C_3)$

$$= 4 \times 11^2 + 3 \times 4^2$$

$$= 532$$

and $S \sim N(468, 532)$

$\therefore P(S > 440) \approx 0.888$

Assumption: The random variables $A_1, A_2, A_3, A_4, C_1, C_2$, and C_3 are independent.

- 10** Let C be the amount of black coffee dispensed and let F be the amount of froth.

Then $C \sim N(120, 7^2)$ and $F \sim N(28, 4.5^2)$

Consider $S = C + F$

$$E(S) = E(C) + E(F)$$

$$= 120 + 28$$

$$= 148 \text{ mL}$$

$$\text{Var}(S) = \text{Var}(C) + \text{Var}(F)$$

$$= 49 + 4.5^2$$

$$= 69.25 \text{ mL}^2$$

$P(U < 135.5) \approx 0.0665 \approx 6.65\%$ which is greater than 1%

$\therefore$ proprietor needs to adjust the machine.

- 11** $X \sim N(-10, 1)$ and $Y \sim N(25, 25)$

a $E(U) = E(3X + 2Y)$

$$= 3E(X) + 2E(Y)$$

$$= 3(-10) + 2(25)$$

$$= 20$$

$$\text{Var}(U) = 9\text{Var}(X) + 4\text{Var}(Y)$$

$$= 9(1) + 4(25)$$

$$= 109$$

$\therefore$ standard deviation of $U \approx 10.4$

b $P(U < 0) \approx 0.0277$ {using technology}

- 12** $S \sim N(280, 4)$ and $L \sim N(575, 16)$

- a** We need to find $P(L < 2S)$, which is $P(L - 2S < 0)$.

If $D = L - 2S$,

$$E(D) = E(L) - 2E(S)$$

$$= 575 - 2 \times 280$$

$$= 15$$

and $\text{Var}(D) = \text{Var}(L) + 4\text{Var}(S)$

$$= 16 + 4 \times 4$$

$$= 32$$

$\therefore D \sim N(15, 32)$

and $P(D < 0) \approx 0.00401$ {using technology}

- b** Now we need to find

$$P(L < S_1 + S_2)$$

$$= P(L - S_1 - S_2 < 0)$$

where $E(L - S_1 - S_2)$

$$= E(L) - E(S_1) - E(S_2)$$

$$= 575 - 280 - 280$$

$$= 15$$

and $\text{Var}(L - S_1 - S_2) = \text{Var}(L) + \text{Var}(S_1) + \text{Var}(S_2)$

$$= 16 + 4 + 4$$

$$= 24$$

Thus $L - S_1 - S_2 \sim N(15, 24)$

and $P(L - S_1 - S_2 < 0) \approx 0.00110$ {using technology}

- 13** **a** $S \sim N(21, 5)$ and $L \sim N(90, 15)$

We need to find $P(L > 5S)$

$$= P(L - 5S > 0)$$

Now $E(L - 5S) = E(L) - 5E(S)$

$$= 90 - 5 \times 21$$

$$= -15$$

and $\text{Var}(L - 5S) = \text{Var}(L) + 25\text{Var}(S)$

$$= 15 + 25 \times 5$$

$$= 140$$

$\therefore L - 5S \sim N(-15, 140)$

and $P(L - 5S > 0) \approx 0.102$ {using technology}

- b** We need to find $P(L > S_1 + S_2 + S_3 + S_4 + S_5)$

$$= P(L - S_1 - S_2 - S_3 - S_4 - S_5 > 0)$$

Now $E(L - S_1 - S_2 - S_3 - S_4 - S_5)$

$$= E(L) - E(S_1) - E(S_2) - E(S_3) - E(S_4) - E(S_5)$$

$$= 90 - 21 - 21 - 21 - 21 - 21$$

$$= -15$$

and $\text{Var}(L - S_1 - S_2 - S_3 - S_4 - S_5)$

$$= \text{Var}(L) + \text{Var}(S_1) + \text{Var}(S_2) + \text{Var}(S_3) + \text{Var}(S_4)$$

$$+ \text{Var}(S_5)$$

$$= 15 + 5 \times 5$$

$$= 40$$

$\therefore L - S_1 - S_2 - S_3 - S_4 - S_5 \sim N(-15, 40)$

and $P(L - S_1 - S_2 - S_3 - S_4 - S_5 > 0)$

$$\approx 0.00885$$

- 14** **a** As $\sum P(x) = 1$ in each distribution, each is a well defined probability distribution.

b $\mu_X = E(X)$

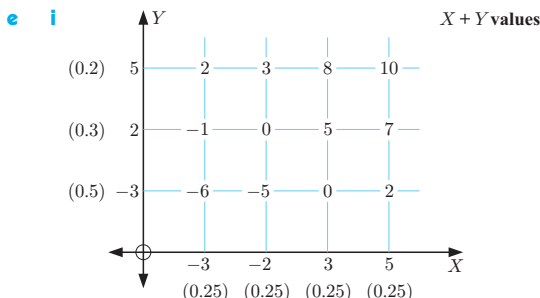
$$= \sum x P(x)$$

$$= -3(0.25) - 2(0.25) + 3(0.25) + 5(0.25)$$

$$= 0.75$$

$$\begin{aligned}
 \text{Var}(X) &= E(X^2) - [E(X)]^2 \\
 &= 9(0.25) + 4(0.25) + 9(0.25) + 25(0.25) - 0.75^2 \\
 &= 47 \times 0.25 - 0.75^2 \\
 &= 11.1875 \text{ and so } \sigma_X \approx 3.34 \\
 \mu_Y &= E(Y) = -3(0.5) + 2(0.3) + 5(0.2) \\
 &= 0.1 \\
 \text{Var}(Y) &= E(Y^2) - [E(Y)]^2 \\
 &= 9(0.5) + 4(0.3) + 25(0.2) - 0.1^2 \\
 &= 10.69 \text{ and so } \sigma_Y \approx 3.27
 \end{aligned}$$

- c With X , the expected win is \$0.75/game. However, it costs \$1 to play each game, so there is an expected loss of \$0.25/game. With Y there is an expected loss of \$0.90/game.
- d As $\sigma_X > \sigma_Y$ we expect greater variation in the results of game X .



$$\begin{aligned}
 P(-6) &= 0.25 \times 0.5 & P(-5) &= 0.25 \times 0.5 \\
 &= 0.125 & &= 0.125 \\
 P(-1) &= 0.25 \times 0.3 & P(0) &= 0.25 \times 0.5 \\
 &= 0.075 & &+ 0.25 \times 0.3 \\
 & & &= 0.200 \\
 P(2) &= 0.25 \times 0.2 & P(3) &= 0.25 \times 0.2 \\
 &+ 0.25 \times 0.5 & &= 0.050 \\
 &= 0.175 \\
 P(5) &= 0.25 \times 0.3 & P(7) &= 0.25 \times 0.3 \\
 &= 0.075 & &= 0.075 \\
 P(8) &= 0.25 \times 0.2 & P(10) &= 0.25 \times 0.2 \\
 &= 0.050 & &= 0.050
 \end{aligned}$$

$X + Y$	-6	-5	-1	0	2
$P(X + Y)$	0.125	0.125	0.075	0.2	0.175

$X + Y$	3	5	7	8	10
$P(X + Y)$	0.05	0.075	0.075	0.05	0.05

- ii $U = X + Y$
 $E(U) = -6(0.125) - 5(0.125) - 1(0.075) + \dots + 10(0.05)$
 $= 0.85$
 $\therefore \mu_U = 0.85$
 $\text{Var}(U) = 36(0.125) + 25(0.125) + 1(0.075) + \dots + 100(0.05) - (0.85)^2$
 $= 21.8775$
 $\therefore \sigma_U \approx 4.68$
- iii With the new game there is an expected loss of \$1 - \$0.85 = \$0.15/game.

EXERCISE B.1

1 a

x	5	10	15	20	25	30
$P(X = x)$	k	k	k	k	k	k

$$6k = 1$$

$$\therefore k = \frac{1}{6}$$

The probability distribution is:

x	5	10	15	20	25	30
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- b $\mu = E(X)$
 $= 5(\frac{1}{6}) + 10(\frac{1}{6}) + 15(\frac{1}{6}) + \dots + 30(\frac{1}{6})$
 $= 17.5$
- c $P(X < \mu) = P(X < 17.5)$
 $= P(X = 5, 10, \text{ or } 15)$
 $= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$
 $= \frac{1}{2}$
- d $\text{Var}(X) = E(X^2) - [E(X)]^2$
 $= 25(\frac{1}{6}) + 100(\frac{1}{6}) + 225(\frac{1}{6}) + 400(\frac{1}{6}) + 625(\frac{1}{6}) + 900(\frac{1}{6}) - 17.5^2$
 $= 72.91666\dots$
 $\therefore \sigma = \sqrt{\text{Var}(X)} \approx 8.54$

- 2 a $P(X = 1) = \frac{3}{4} = p$
 $P(X = 0) = \frac{1}{4} = 1 - p$
 $\therefore X \sim B(1, \frac{3}{4})$

b $F(x) = \sum_{k=0}^x p^k (1-p)^{1-k}$
 $\therefore F(0) = (\frac{3}{4})^0 (\frac{1}{4})^1 = \frac{1}{4}$
 $F(1) = (\frac{3}{4})^0 (\frac{1}{4})^1 + (\frac{3}{4})^1 (\frac{1}{4})^0 = 1$

Interpretation:

$F(0)$ is the probability of 'no reds'

$F(1)$ is the probability of 'at most one red'

'at most one red' is 0 reds or 1 red which is a certain event.

- 3 $X \sim B(7, p)$ where $p < 0.5$.
 $P(X = 4) = \binom{7}{4} p^4 (1-p)^3 = 0.09724$
 $\therefore 35p^4 (1-p)^3 = 0.09724$
 $\therefore p^4 (1-p)^3 \approx 0.00277829$
 $\Rightarrow p \approx 0.29999943 \text{ or } 0.81548136$
 But $p < 0.5$, so $p \approx 0.300$
 and $P(X = 2) = \binom{7}{2} (0.3)^2 (0.7)^5$
 ≈ 0.318

- 4 $p = 0.35$ is the probability of rain on an August day.
 $X \sim B(7, 0.35)$
- a $P(X = 3) = \binom{7}{3} (0.35)^3 (0.65)^4$
 ≈ 0.268
- b $P(X \geq 3) = 1 - P(X \leq 2)$
 ≈ 0.468
- c $P(X \leq 3) \approx 0.800$

d Days of rain:	D ₁	D ₂	D ₃
	1st	2nd	3rd
	2nd	3rd	4th
	3rd	4th	5th
	4th	5th	6th
	5th	6th	7th

$$\begin{aligned} \therefore P(\text{rains on exactly 3 days in succession}) \\ = 5 \times (0.35)^3 (0.65)^4 \\ \approx 0.0383 \end{aligned}$$

Assumptions:

Rain falling on any day is independent of rain falling on any other day.

- 5 Let X be the number of red pens selected.

Due to the very large number of pens, the number of reds selected in n attempts is approximately $X \sim B(n, 0.2)$

$$\text{Now } P(X \geq 1) > 0.9$$

$$\therefore P(X = 0) < 0.1$$

$$\therefore \binom{n}{0} (0.2)^0 (0.8)^n < 0.1$$

$$\therefore (0.8)^n < 0.1$$

$$\therefore n \log(0.8) < \log(0.1)$$

$$\therefore n > \frac{\log(0.1)}{\log(0.8)} = 10.318\dots$$

$$\therefore \text{least } n \text{ is } n = 11.$$

We are assuming the binomial model even though it is not strictly binomial.

- 6 a Let X be the number of cells failing in one year.

$$X \sim B(15, 0.7)$$

$$\text{i } P(X = 15) \approx 0.00475$$

$$\begin{aligned} \text{ii } P(\text{still operating}) &= P(X \leq 14) \\ &\approx 1 - 0.00475 \\ &\approx 0.99525 \\ &\approx 0.995 \end{aligned}$$

- b $P(\text{still operating with } n \text{ cells})$

$$= P(X \leq n - 1)$$

$$= 1 - P(X = n)$$

$$= 1 - (0.7)^n$$

- c We need to find the smallest n such that

$$1 - (0.7)^n \geq 0.98$$

$$\therefore (0.7)^n \leq 0.02$$

$$\therefore n \log(0.7) \leq \log(0.02)$$

$$\therefore n \geq \frac{\log(0.02)}{\log(0.7)}$$

$$\therefore n \geq 10.968\dots$$

$$\therefore \text{least } n \text{ is } n = 11.$$

- 7 Let X be the number of letters addressed to the Accounts Department.

$$X \sim B(20, 0.7)$$

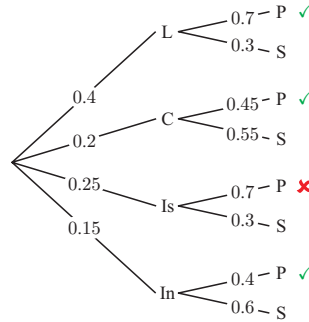
$$\begin{aligned} \text{a } P(X \geq 11) &= 1 - P(X \leq 10) \\ &\approx 0.952 \end{aligned}$$

- b $Y \sim B(70, 0.7)$

$$\mu = np = 70 \times 0.7 = 49 \text{ letters}$$

$$\begin{aligned} \sigma &= \sqrt{np(1-p)} \\ &= \sqrt{70 \times 0.7 \times 0.3} \\ &\approx 3.83 \end{aligned}$$

8



- a $P(\text{Is} | P)$

$$\begin{aligned} &= \frac{P(P | \text{Is})P(\text{Is})}{P(P)} \quad \{\text{Bayes Theorem}\} \\ &= \frac{(0.25)(0.7)}{(0.4)(0.7) + (0.2)(0.45) + (0.25)(0.7) + (0.15)(0.4)} \\ &\approx 0.289 \end{aligned}$$

- b For one parcel,

$P(\text{leaves state} | S)$

$$\begin{aligned} &= \frac{P(S | \text{leaves state})P(\text{leaves state})}{P(S)} \\ &= \frac{(0.25)(0.3) + (0.15)(0.6)}{(0.4)(0.3) + (0.2)(0.55) + (0.25)(0.3) + (0.15)(0.6)} \\ &\approx 0.41772 \end{aligned}$$

If X is the number of parcels selected then

$$X \sim B(2, 0.41772) \text{ and } P(X = 1) \approx 0.486.$$

Assumption: The events are independent.

- 9 a There are 7 multiples of 7 which are < 50 .

$$\therefore P(\text{multiple of } 7) = \frac{7}{50} = 0.14$$

- b Let X = number of multiples of 7 obtained in 500 spins.

Then $X \sim B(500, 0.14)$

Now 15% of 500 = 75 and

$$\begin{aligned} P(X > 75) &= 1 - P(X \leq 75) \\ &\approx 0.237 \quad \{\text{using technology}\} \end{aligned}$$

- c i $E(X) = np$

$$= 500 \times 0.14$$

$$= 70$$

$\therefore$ the school expects to make in 500 spins

$$(500 - 70) \times \$20 - 70 \times \$100 = \$1600$$

- ii The school loses if

$$(500 - X)20 - 100X < 0$$

$$\therefore 500 - X - 5X < 0$$

$$\therefore 6X > 500$$

$$\therefore X > 83\frac{1}{3}$$

$$\text{and } P(X > 83\frac{1}{3}) = 1 - P(X \leq 83)$$

$$\approx 0.0435$$

EXERCISE B.2

- 1 $X \sim \text{Geo}(0.25)$

$$\text{a } P(X = 4)$$

$$= p(1-p)^3$$

$$= 0.25 \times (0.75)^3$$

$$\approx 0.105$$

$$\text{b } P(X \leq 2)$$

$$= P(X = 1 \text{ or } 2)$$

$$= p + p(1-p)$$

$$= 0.4375$$

- c $P(X > 3)$
 $= 1 - P(X \leq 3)$
 $= 1 - P(X = 1, 2, \text{ or } 3)$
 $= 1 - [p + p(1-p) + p(1-p)^2]$
 ≈ 0.422
- 2 $X \sim \text{Geo}(p)$, then $P(X = x) = p(1-p)^{x-1}$
 for $x = 1, 2, 3, 4, 5, \dots$
 Now

$$\sum_{i=1}^{\infty} P(X = i) = P(X = 1) + P(X = 2) + P(X = 3) + \dots$$

$$= p + p(1-p) + p(1-p)^2 + p(1-p)^3 + \dots$$

$$= p[1 + (1-p) + (1-p)^2 + (1-p)^3 + \dots]$$

$$= p \left[\frac{1}{1 - (1-p)} \right]$$
 {as $1 + (1-p) + (1-p)^2 + \dots$ is an infinite GS
 with $u_1 = 1$ and $r = 1-p$ with $0 < r < 1$.}

$$= p \left[\frac{1}{p} \right]$$

$$= 1$$

 $\therefore$ the probability distribution is well defined.
- 3 a $X \sim \text{Geo}(0.29)$ Using technology,
 $\therefore P(X = 4) \approx 0.104$
 b $Y \sim \text{NB}(3, 0.29)$
 $\therefore P(X = 7) = \binom{6}{2} (0.29)^3 (0.71)^4$
 ≈ 0.0930
- 4 $X \sim \text{Geo}(p)$
 $P(X = 3) = p(1-p)^2 = 0.023987$
 Solving using technology gives $p = 0.83$ {as $p > 0.5$ }
 and $P(X \geq 3) = 1 - P(X \leq 2)$
 $= 0.0289$ {using technology}
- 5 a $X = \text{number of games needed for Eva to win 3 games}$
 Then $X \sim \text{NB}(3, 0.35)$
 b $P(\text{Eva beats Paul 3 games to 1})$
 $= P(X = 4)$
 $= \binom{3}{2} (0.35)^3 (0.65)^1$
 ≈ 0.0836
 c $P(\text{Eva beats Paul in a match})$
 $= P(X = 3, 4, \text{ or } 5)$
 $= \binom{2}{2} (0.35)^3 (0.65)^0 + \binom{3}{2} (0.35)^3 (0.65)^1$
 $+ \binom{4}{2} (0.35)^3 (0.65)^2$
 ≈ 0.235
- 6 a $X \sim \text{Geo}(0.72)$
 $\therefore P(X = 5) \approx 0.00443$
 b $Y \sim \text{NB}(4, 0.72)$
 $\therefore P(X = 12) = \binom{11}{3} (0.72)^4 (0.28)^8$
 ≈ 0.00168
- 7 $X \sim \text{Geo}(0.15)$
 a $P(\text{first snow on Nov 15})$
 $= P(X = 15)$
 ≈ 0.0154

- b $P(\text{snow falls on or before } n \text{ days})$
 $= 1 - P(\text{snow does not fall in } n \text{ days})$
 $= 1 - (0.85)^n$
 So, we need to solve
 $1 - (0.85)^n > 0.85$
 $\therefore (0.85)^n < 0.15$
 Thus, $n \log(0.85) < \log(0.15)$
 $\therefore n > \frac{\log(0.15)}{\log(0.85)}$ { $\log 0.85 < 0$ }
 $\therefore n > 11.673\dots$
 $\therefore$ least n is $n = 12$
 $\therefore$ must book for Dec 12.

8 a

Difference table

	6	5	4	3	2	1	0
	5	4	3	2	1	0	1
Die 2	4	3	2	1	0	1	2
	3	2	1	0	1	2	3
	2	1	0	1	2	3	4
	1	0	1	2	3	4	5
		1	2	3	4	5	6

Die 1

$$P(\text{difference is no more than 3}) = \frac{30}{36} = \frac{5}{6}$$

- b $X \sim G(\frac{5}{6})$
 $\therefore P(\text{player 1 is first to start on 2nd roll})$
 $= P(X = 5)$ {all 4 players fail on 1st attempt}
 ≈ 0.000643

EXERCISE B.3

- 1 a $P(X = x) = \frac{m^x e^{-m}}{x!}$ for $x = 0, 1, 2, 3, 4, \dots$
 $P(X = 2) = P(X = 0) + 2P(X = 1)$
 $\therefore \frac{m^2 e^{-m}}{2!} = \frac{e^{-m}}{0!} + \frac{2m e^{-m}}{1!}$
 $\therefore \frac{m^2}{2} = 1 + 2m$
 $\therefore m^2 = 4m + 2$
 $\therefore m^2 - 4m - 2 = 0$
 $\therefore m = \frac{4 \pm \sqrt{16 - 4(1)(-2)}}{2}$
 $\therefore m = 2 \pm \sqrt{6}$
 But $m > 0$, so $m = 2 + \sqrt{6} \approx 4.44948\dots$
 $\therefore \mu \approx 4.45$
- b $P(1 \leq X \leq 5)$
 $= P(X \leq 5) - P(X = 0)$
 $\approx 0.71153 - 0.01168$
 ≈ 0.700
- 2 a X is a Poisson random variable as the average number of phone calls to the police per hour is constant. We assume that the average number of calls each hour is constant.
 b i Since $2\text{Var}(X) = [E(X)]^2 - 15$,
 $2m = m^2 - 15$
 $\therefore m^2 - 2m - 15 = 0$
 $\therefore (m+3)(m-5) = 0$
 $\therefore m = 5$ {as $m > 0$ }
 $\therefore$ the mean is 5 calls/hour.

ii Thus $X \sim \text{Po}(5)$ and $P(X \leq 3) \approx 0.265$.

3 a i The probability of a fault in a 50 m length $= \frac{50}{1000}$
 $= 0.05$

$$\therefore X \sim \text{Po}(0.05)$$

$$\therefore P(X = 0) \approx 0.951$$

ii $P(\text{at most 2 faults in 50 m})$

$$= P(X \leq 2)$$

$$\approx 0.99998 \text{ which is } \approx 1$$

b $P(X \leq 1) \approx 0.9988$ which is > 0.995
 So, the chain is considered safe.

4 a Let X = number of internal calls, and
 Y = number of external calls

$$\therefore X \sim \text{Po}(\frac{5}{4}) \text{ and } Y \sim \text{Po}(\frac{10}{6}), \text{ for 5 minutes.}$$

The total number of calls each five minute session is $X + Y$
 where $E(X + Y) = E(X) + E(Y)$

$$= \frac{5}{4} + \frac{10}{6}$$

$$\approx 2.917$$

$$\text{and } \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\approx 2.917 \text{ also}$$

$$\therefore X + Y \sim \text{Po}(2.917)$$

{assuming X and Y are independent random variables}

$$\therefore P(X + Y = 3) \approx 0.224$$

b As $E(X + Y) \approx 2.917$, the receptionist should expect
 3 calls each 5 minutes.

c i $P(X + Y > 5) = 1 - P(X + Y \leq 5)$
 ≈ 0.0758

ii 5 calls in 20 min = x calls in 7 min

$$\Rightarrow x = \frac{7}{4}$$

10 calls in 30 min = y calls in 7 min

$$\Rightarrow y = \frac{7}{3}$$

$$\therefore E(X + Y) = \text{Var}(X + Y) = \frac{7}{4} + \frac{7}{3} \approx 4.083$$

$$\text{and } P(X + Y > 5) = 1 - P(X + Y \leq 5)$$

$$\approx 0.228$$

$$\{\text{as } X + Y \sim \text{Po}(4.083)\}$$

5 a $X \sim \text{Po}(m_1)$ and $Y \sim \text{Po}(m_2)$ are two independent
 random variables.

$$\therefore P(X = x) = \frac{m_1^x e^{-m_1}}{x!} \text{ and}$$

$$P(Y = y) = \frac{m_2^y e^{-m_2}}{y!}$$

Since X and Y are independent

$$P(X = x \text{ and } Y = y)$$

$$= P(X = x) \times P(Y = y)$$

Now $P(X + Y = k)$

$$= \sum_{i=0}^k [P(X = i \text{ and } Y = k - i)]$$

$$= \sum_{i=0}^k [P(X = i) \times P(Y = k - i)]$$

$$= \sum_{i=0}^k \frac{m_1^i e^{-m_1}}{i!} \times \frac{m_2^{k-i} e^{-m_2}}{(k-i)!}$$

$$= \frac{e^{-m_1-m_2}}{k!} \sum_{i=0}^k \frac{k!}{i!(k-i)!} m_1^i m_2^{k-i}$$

$$= \frac{e^{-(m_1+m_2)}}{k!} \times (m_1 + m_2)^k$$

{Binomial theorem}

Thus, $X + Y$ is Poisson with mean $m_1 + m_2$.

b P_n is "If $X_1, X_2, X_3, \dots, X_n$ are independent Poisson
 random variables with corresponding means $m_1, m_2, m_3,$
 $\dots, m_n$ then $X_1 + X_2 + \dots + X_n$ is a Poisson random
 variable with mean $m_1 + m_2 + m_3 + \dots + m_n$ "

Proof by induction:

$$\text{If } n = 1, P(X = x) = \frac{m_1^x e^{-m_1}}{x!}.$$

$\therefore P_1$ is true.

If P_j is true then $X_1 + X_2 + \dots + X_j$

has $P(X_1 + X_2 + \dots + X_j = l)$ has PDF

$$\frac{(m_1 + m_2 + \dots + m_j)^l e^{-(m_1+m_2+\dots+m_j)}}{l!}$$

$\therefore P(X_1 + X_2 + \dots + X_j + X_{j+1} = l)$ has PDF

$$\sum_{i=0}^l P(X_1 + X_2 + \dots + X_j = i \text{ and } X_{j+1} = l - i)$$

$$= \sum_{i=0}^l \left[\frac{(m_1 + m_2 + \dots + m_j)^i e^{-(m_1+m_2+\dots+m_j)}}{i!} \right.$$

$$\times \left. \frac{m_{j+1}^{l-i} e^{-m_{j+1}}}{(l-i)!} \right]$$

$$= \frac{e^{-(m_1+m_2+\dots+m_j+m_{j+1})}}{l!} \times$$

$$\sum_{i=0}^l (m_1 + m_2 + \dots + m_j)^i \times m_{j+1}^{l-i} \times \frac{l!}{i!(l-i)!}$$

$$= \frac{e^{-(m_1+m_2+\dots+m_{j+1})}}{l!}$$

$$\times (m_1 + m_2 + \dots + m_j + m_{j+1})^l$$

Thus P_{j+1} is true whenever P_j is true, and P_1 is true.

$\Rightarrow P_n$ is true. {Principle of mathematical induction}

EXERCISE B.4

$$1 \text{ a } \mu = \frac{1}{p}$$

$$= \frac{1}{0.333}$$

$$\approx 3.00$$

$$b \text{ Var}(X) = \frac{q}{p^2}$$

$$= \frac{0.667}{(0.333)^2}$$

$$\approx 6.015 \dots$$

$$\therefore \sigma \approx 2.45$$

2 a $X \sim \text{Geo}(0.25)$

$$\therefore \mu = \frac{1}{p}$$

$$= \frac{1}{0.25}$$

$$= 4 \text{ throws}$$

b $Y \sim \text{NB}(2, 0.25)$

$$\therefore \mu = \frac{r}{p}$$

$$= \frac{2}{0.25}$$

$$= 8 \text{ throws}$$

3 $X \sim \text{Geo}(0.05)$

$$a \mu = \frac{1}{p}$$

$$= \frac{1}{0.05}$$

$$= 20 \text{ throws}$$

$$b \text{ Var}(X) = \frac{q}{p^2}$$

$$= \frac{0.95}{(0.05)^2}$$

$$\therefore \sigma \approx \sqrt{380}$$

$$\therefore \sigma \approx 19.5 \text{ throws}$$

4 $X \sim \text{DU}(40)$

$$\begin{aligned} \text{a } \mu &= \frac{n+1}{2} \\ &= \frac{41}{2} \\ &= 20.5 \end{aligned}$$

$$\begin{aligned} \text{b } \text{Var}(X) &= \frac{n^2 - 1}{12} \\ &= \frac{40^2 - 1}{12} \\ &= 133.25 \\ \therefore \sigma &= \sqrt{133.25} \\ &\approx 11.5 \end{aligned}$$

5 **A** $X \sim \text{Po}(6)$. $P(X = 3) \approx 0.0892$

B $X \sim \text{Po}(1)$. $P(X = 1) \approx 0.3679$

C $X \sim \text{Po}(24)$. $P(X = 17) \approx 0.0308$

As **B** has the highest probability it is the most likely to occur.

6 Let X = Yankees beat Redsox in a game

$X \sim \text{NB}(4, 0.47)$

$$\begin{aligned} \text{a } P(X = 5) &= \binom{4}{3} (0.47)^4 (0.53)^1 \\ &\approx 0.103 \\ \text{b } P(X = 7) &= \binom{6}{3} (0.47)^4 (0.53)^3 \\ &\approx 0.145 \end{aligned}$$

$$\begin{aligned} \text{c } P(\text{Redsox win}) &= P(X = 0, 1, 2, \text{ or } 3) \\ &= 1 - P(X = 4, 5, 6, \text{ or } 7) \\ &= 1 - \left[\binom{3}{3} (0.47)^4 (0.53)^0 + \binom{4}{3} (0.47)^4 (0.53)^1 \right. \\ &\quad \left. + \binom{5}{3} (0.47)^4 (0.53)^2 + \binom{6}{3} (0.47)^4 (0.53)^3 \right] \\ &\approx 0.565 \end{aligned}$$

d $X \sim \text{NB}(4, 0.53)$

$$\text{and } E(X) = \frac{r}{p} = \frac{4}{0.53} \approx 7.547 \text{ games}$$

This is the average number of games it would take them to win without restriction, i.e., by playing as many games as they need. However, in a World Series, no more than 7 games will be played (assuming no draws) to decide the winner.

7 **a** If X is the number of attempts needed then $X \sim \text{Geo}(0.62)$. This assumes that attempts are independent and the probability of getting through remains constant.

$$\begin{aligned} \text{b } P(X \geq 3) &= 1 - P(X \leq 2) \\ &\approx 0.144 \end{aligned}$$

$$\begin{aligned} \text{c } \mu &= \frac{1}{p} \\ &= \frac{1}{0.62} \\ &\approx 1.61 \\ \sigma &= \sqrt{\frac{1-p}{p^2}} \\ &= \sqrt{\frac{1-0.62}{0.62^2}} \\ &\approx 0.994 \end{aligned}$$

8 Let X = the number who do not arrive.

a Then $X \sim (255, 0.0375)$

$$\begin{aligned} \text{b } P(\text{more than 250 arrive}) &= P(X \leq 4) \\ &\approx 0.0362 \end{aligned}$$

$$\begin{aligned} \text{c } P(\text{there are empty seats}) &= P(X \geq 6) \\ &= 1 - P(X \leq 5) \\ &\approx 0.918 \end{aligned}$$

$$\begin{aligned} \text{d } \text{i } \mu &= np \\ &= 255 \times 0.0375 \\ &\approx 9.56 \\ \text{ii } \sigma^2 &= np(1-p) \\ &= 9.20 \end{aligned}$$

e As $\mu \approx \sigma^2$, $n \geq 50$ and $p \leq 0.1$ a Poisson distribution could be used to approximate the binomial distribution where $X \sim \text{Po}(10)$.

$$\text{i } P(X \leq 4) \approx 0.0293$$

$$\begin{aligned} \text{ii } P(X \geq 6) &= 1 - P(X \leq 5) \\ &\approx 0.933 \end{aligned}$$

f The Poisson approximation is reasonably good.

9 **a** X = a return from playing the game

$$\begin{aligned} &= -€14.90, -€14.80, -€14.70, -€14.60, -€14.50, \\ &\quad -€14.40, -€14.30, €0, €15, €85 \end{aligned}$$

$$\begin{aligned} \text{b } E(X) &= \sum p_i x_i \\ &= \frac{1}{10}(-14.90) + \frac{1}{10}(-14.80) + \frac{1}{10}(-14.70) \\ &\quad + \dots + \frac{1}{10}(85) \text{ Euro} \\ &= -€0.22 \end{aligned}$$

$$\begin{aligned} \text{and } \text{Var}(X) &= \sum x_i^2 p - [E(X)]^2 \\ &= (-14.90)^2(0.1) + (-14.80)^2(0.1) \\ &\quad + \dots + (85)^2(0.1) - (-0.22)^2 \\ &\approx 894 \end{aligned}$$

c If $X \sim \text{DU}(10)$, it assumes that X has values 1, 2, 3, 4, ..., 10 which is not the case here.

d **i** For a game costing €15 the expected loss is 22 cents
 $\therefore$ for a game costing €14.80 the expected loss is 2 cents
 $\therefore$ the smallest amount is €14.80.

ii For each game $E(X) = -€1.22$
 $\therefore$ for 1000 games they would expect to make
 $1000 \times €1.22 = €1220$

10 **a** $X \sim \text{Geo}(\frac{1}{8})$

Assumptions:

- each call is made with probability $\frac{1}{8}$ of success
- calls are independent of each other

$$\begin{aligned} \text{b } E(X) &= \frac{1}{p} = 8, \quad \text{Var}(X) = \frac{1-p}{p^2} \\ &= \frac{\frac{7}{8}}{\frac{1}{64}} \\ &= 56 \end{aligned}$$

$$\therefore \mu = 8 \text{ and } \sigma \approx 7.48$$

$$\begin{aligned} \text{c } P(X < 5) &= P(X \leq 4) \\ &\approx 0.414 \end{aligned}$$

11 **a** T = number of wrong numbers dialled in a typical week
 $\therefore T \sim \text{B}(75, 0.005)$

$$\text{b } \text{i } P(T = 0) \approx 0.687$$

$$\begin{aligned} \text{ii } P(T > 2) &= 1 - P(T \leq 2) \\ &\approx 0.00646 \end{aligned}$$

$$\text{c } E(T) = np = 0.375$$

$$\text{Var}(T) = np(1-p) = 0.373$$

The mean and variance are almost the same which suggests that T can be approximated using a Poisson distribution.

d Using $T \sim \text{Po}(0.375)$

$$\text{i } P(T = 0) \approx 0.687$$

$$\begin{aligned} \text{ii } P(T > 2) &= 1 - P(T \leq 2) \\ &\approx 0.00665 \end{aligned}$$

- e** Both results are very close verifying that for large n and small p , the binomial distribution can be approximated by the Poisson distribution with the same mean i.e., $X \sim \text{Po}(np)$.

12 a i As $0 < q < 1$,
 $1 + q + q^2 + q^3 + \dots$ has sum to infinity $\frac{1}{1-q}$
 {sum of an infinite geometric series}

ii Thus $1 + q + q^2 + q^3 + q^4 + \dots = (1-q)^{-1}$
 Differentiating both sides with respect to q gives:
 $1 + 2q + 3q^2 + 4q^3 + \dots = -(1-q)^{-2} \times (-1)$

$$= \frac{1}{(1-q)^2}$$

$$\Rightarrow \sum_{x=1}^{\infty} xq^{x-1} = \frac{1}{(1-q)^2} \text{ for } 0 < q < 1$$

b $X \sim \text{Geo}(p)$

$$\begin{aligned} \therefore P(X=x) &= p(1-p)^{x-1} \\ &= pq^{x-1} \text{ where } q = 1-p \end{aligned}$$

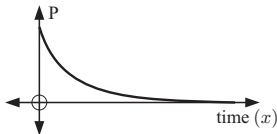
$$\begin{aligned} \text{Now } E(X) &= \sum_{x=1}^{\infty} xP(X=x) \\ &= \sum_{x=1}^{\infty} xpq^{x-1} \\ &= p \sum_{x=1}^{\infty} xq^{x-1} \\ &= p \times \frac{1}{(1-q)^2} \quad \{\text{from a ii}\} \\ &= p \times \frac{1}{p^2} \\ &= \frac{1}{p} \end{aligned}$$

EXERCISE C

1 $T \sim U(-\pi, \pi)$

$$\begin{aligned} \mu &= \frac{a+b}{2} \\ &= \frac{-\pi + \pi}{2} \\ &= 0 \end{aligned} \quad \begin{aligned} \sigma &= \sqrt{\frac{(a-b)^2}{12}} \\ &= \sqrt{\frac{(2\pi)^2}{12}} \\ &= \sqrt{\frac{\pi^2}{3}} \\ &= \frac{\pi}{\sqrt{3}} \end{aligned}$$

- 2 a** The best chance of getting a ticket is as soon as possible after release. As time goes by it gets increasingly difficult and very quickly almost impossible.
 The distribution has the shape shown, and the variable X is continuous.



b As the median is 10,

$$\begin{aligned} \int_0^{10} f(x) dx &= 0.5 \\ \therefore \int_0^{10} \lambda e^{-\lambda x} dx &= 0.5 \\ \therefore \lambda \left[\frac{1}{-\lambda} e^{-\lambda x} \right]_0^{10} &= 0.5 \\ \therefore (-e^{-10\lambda}) - (-e^0) &= 0.5 \\ \therefore e^{-10\lambda} &= 0.5 \\ \therefore -10\lambda = \ln \frac{1}{2} = -\ln 2 \\ \therefore \lambda &= \frac{\ln 2}{10} \approx 0.0693 \end{aligned}$$

c $P(\text{seat purchased after 3 days})$

$$\begin{aligned} &= P(X \geq 72) \\ &= 1 - P(X < 72) \\ &= 1 - \int_0^{72} 0.069315e^{-0.069315x} dx \\ &= 0.00680 \end{aligned}$$

d $E(X) = \frac{1}{\lambda} \approx 14.4$ hours

The average time to buy a ticket is ≈ 14.4 hours.

3 $X \sim N(\mu, \sigma^2)$

If $P(X > 13) = 0.4529$ then

$$\begin{aligned} P\left(\frac{X-\mu}{\sigma} < \frac{13-\mu}{\sigma}\right) &= 0.5471 \\ \therefore P\left(Z < \frac{13-\mu}{\sigma}\right) &= 0.5471 \\ \therefore \frac{13-\mu}{\sigma} &= \text{invNorm}(0.5471) \\ \therefore 13-\mu &\approx 0.11834 \dots (1) \end{aligned}$$

If $P(X > 28) = 0.1573$ then

$$\begin{aligned} P\left(Z < \frac{28-\mu}{\sigma}\right) &= 0.8427 \\ \therefore \frac{28-\mu}{\sigma} &\approx 1.00562 \\ \therefore 28-\mu &\approx 1.00562\sigma \dots (2) \end{aligned}$$

Solving (1) and (2) simultaneously gives $\mu \approx 11.0$ and $\sigma \approx 16.9$

4 a We require $\int_0^k (6-18x) dx = 1$

$$\begin{aligned} \therefore [6x - 9x^2]_0^k &= 1 \\ \therefore 6k - 9k^2 &= 1 \\ \therefore 9k^2 - 6k + 1 &= 0 \\ \therefore (2k-1)^2 &= 0 \\ \therefore k &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} \text{b } \mu = E(X) &= \int_0^{\frac{1}{3}} x f(x) dx \\ &= \int_0^{\frac{1}{3}} (6x - 18x^2) dx \\ &= \frac{1}{9} \end{aligned}$$

$$\begin{aligned}\sigma^2 &= E(X^2) - [E(X)]^2 \\ &= \int_0^{\frac{1}{3}} x^2 f(x) dx - \left(\frac{1}{9}\right)^2 \\ &= \int_0^{\frac{1}{3}} (6x^2 - 18x^3) dx - \frac{1}{81} \\ &\approx 0.0061728\end{aligned}$$

$$\therefore \sigma \approx 0.0786$$

- 5 a X = number who support the Environment Party
 $X \sim B(180, 0.41)$

$$\begin{aligned}\text{b } E(X) &= np & \text{Var}(X) &= np(1-p) \\ &= 180 \times 0.41 & &= 180 \times 0.41 \times 0.59 \\ &= 73.8 & &\approx 43.5\end{aligned}$$

$$\begin{aligned}\text{c } P(X \geq 58) &= 1 - P(X \leq 57) \\ &\approx 0.994\end{aligned}$$

- d As $np = 73.8 > 5$ and $n(1-p) = 106.2 > 5$ we can approximate this binomial distribution to the normal distribution and $X \sim N(73.8, 43.5)$

$$\begin{aligned}P(X \geq 58) &= P(X^* \geq 57.5) \\ &\approx 0.993\end{aligned}$$

- 6 a X = number of mistakes made on one page
 $X \sim \text{Po}(2.5)$ is the discrete random variable.
 Consider $Y = X_1 + X_2 + X_3 + X_4 + \dots + X_{52}$ where X_i are assumed to be independent.

$$\begin{aligned}E(Y) &= E(X_1) + E(X_2) + \dots + E(X_{52}) \\ &= 52 \times 2.5 \\ &= 130\end{aligned}$$

$$\begin{aligned}\text{and } \text{Var}(Y) &= \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_{52}) \\ &= 52 \times 2.5 \quad \{\text{assumption}\} \\ &= 130 \text{ also.}\end{aligned}$$

$$\text{So, } Y \sim \text{Po}(130)$$

$$\begin{aligned}\text{b i } P(X > 2) & & \text{ii } P(Y > 104) \\ &= 1 - P(X \leq 2) &= 1 - P(Y \leq 104) \\ &\approx 0.456 &\approx 0.989\end{aligned}$$

$$\begin{aligned}\text{c } X &\sim N(2.5, 2.5) \\ \therefore P(X > 2) &= P(X^* \geq 2.5) \\ &\approx 0.500\end{aligned}$$

$$Y \sim N(130, 130)$$

$$\begin{aligned}\therefore P(Y > 104) &= P(Y^* > 104.5) \\ &\approx 0.987\end{aligned}$$

The approximation for X is poor, but that for Y is very good. This is probably due to the fact that $2.5 = \lambda$ is not large enough.

- 7 X is a uniform continuous random variable.

$$\text{a } \int_1^k f(x) dx = 1$$

$$\therefore \left[\frac{2}{5}x\right]_1^k = 1$$

$$\therefore \frac{2k}{5} - \frac{2}{5} = 1$$

$$\therefore 2k - 2 = 5$$

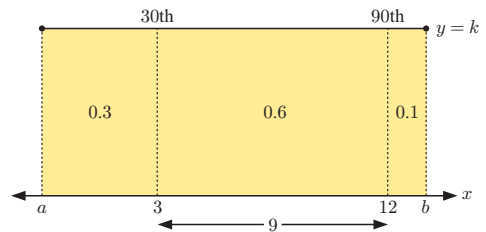
$$\therefore k = 3.5$$

Thus $X \sim U(1, 3.5)$.

$$\begin{aligned}\text{b } P(1.7 \leq X \leq 3.2) &= \int_{1.7}^{3.2} \frac{2}{5} dx \\ &= \left[\frac{2}{5}x\right]_{1.7}^{3.2} \\ &= 0.6\end{aligned}$$

$$\begin{aligned}\text{c } E(X) &= \frac{a+b}{2} & \text{Var}(X) &= \frac{(b-a)^2}{12} \\ &= \frac{1+3.5}{2} & &= \frac{(3.5-1)^2}{12} \\ &= 2.25 & &\approx 0.521\end{aligned}$$

- 8 a Suppose the PDF of X is $f(x) = k$.



$$\text{Now } 9k = 0.6, \text{ so } k = \frac{0.6}{9} = \frac{1}{15}$$

$$\text{Also } (3-a)k = 0.3, \text{ so } \frac{3-a}{15} = 0.3$$

$$\therefore 3-a = 4.5$$

$$\therefore a = -1.5$$

$$\text{and } (3-a)k = 0.3, \text{ so } \frac{3-a}{15} = 0.3$$

$$\therefore b = 13.5$$

$$\text{b } f(x) = \frac{1}{15}, -1.5 \leq x \leq 13.5$$

$$\text{c } P(5 < X < 9) = \frac{4}{15}$$

$$\begin{aligned}\text{d CDF, } F(X) &= \frac{x-a}{b-a} = \frac{x+1.5}{15} \\ \therefore F(X) &= \begin{cases} \frac{2x+3}{30}, & -1.5 < x < 13.5 \\ 0 & \text{elsewhere} \end{cases}\end{aligned}$$

- 9 a $T \sim N(7, 36) \therefore \sigma = 6$

$$\begin{aligned}P(|T-6| < 2.3) &= P(-2.3 < T-6 < 2.3) \\ &= P(3.7 < T < 8.3) \\ &\approx 0.295\end{aligned}$$

- b $X \sim B(4, 0.29461)$ where X is the number of times T lies in the interval $3.7 < T < 8.3$ and $P(X=2) \approx 0.259$

- 10 a $f(x) = \frac{e^{-\frac{1}{2}x}}{2} = 0.5e^{-0.5x}$ which is the PDF of an exponential random variable with $\lambda = 0.5$ i.e., $X \sim \text{Exp}(0.5)$

$$\begin{aligned}\text{b i } \mu_X &= E(X) & \text{ii } \text{Var}(X) \\ &= \frac{1}{\lambda} &= \frac{1}{\lambda^2} \\ &= 2 &= 4 \\ & &\therefore \sigma_X = 2\end{aligned}$$

iii If the median is m , say

$$\begin{aligned}\int_0^m f(x) dx &= 0.5 \\ \therefore 0.5 \int_0^m e^{-0.5x} dx &= 0.5 \\ \therefore \int_0^m e^{-0.5x} dx &= 1 \\ \therefore \left[\frac{1}{-0.5} e^{-0.5x} \right]_0^m &= 1 \\ \therefore -2e^{-\frac{m}{2}} + 2e^0 &= 1 \\ \therefore e^{-\frac{m}{2}} &= \frac{1}{2} \\ \therefore -\frac{m}{2} &= \ln \frac{1}{2} = -\ln 2 \\ \therefore m &= 2 \ln 2 = \ln 4\end{aligned}$$

iv We need to find a , say such that

$$\begin{aligned}\int_0^a 0.5e^{-0.5x} dx &= 0.9 \\ \therefore \int_0^a e^{-0.5x} dx &= 1.8 \\ \therefore \left[\frac{1}{-0.5} e^{-0.5x} \right]_0^a &= 1.8 \\ \therefore e^{-0.5a} - 1 &= -0.9 \\ \therefore e^{-0.5a} &= 0.1 \\ \therefore e^{0.5a} &= 10 \\ \therefore 0.5a &= \ln 10 \\ \therefore a &= 2 \ln 10 = \ln 100\end{aligned}$$

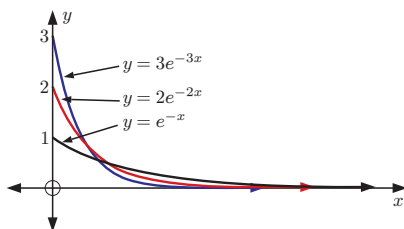
c When $X \sim \text{Exp}(\lambda)$, the CDF of X is

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}$$

$$\therefore P(X \leq x) = 1 - e^{-\frac{x}{2}}$$

$$\begin{array}{ll} \text{i} & P(X \leq 1) \\ &= 1 - e^{-0.5} \\ &\approx 0.3935 \end{array} \quad \begin{array}{ll} \text{ii} & P(0.4 \leq X \leq 2) \\ &= P(X \leq 2) - P(X \leq 0.4) \\ &= (1 - e^{-1}) - (1 - e^{-0.2}) \\ &= e^{-0.2} - e^{-1} \\ &\approx 0.4509 \end{array}$$

11 a $f(x) = ae^{-ax}$ for $a = 1, 2, 3$



$$\begin{aligned}\text{b} \quad \int_0^\infty ae^{-ax} dx &= \lim_{t \rightarrow \infty} \int_0^t ae^{-ax} dx \\ &= \lim_{t \rightarrow \infty} \left[a \left(\frac{1}{-a} \right) e^{-ax} \right]_0^t \\ &= \lim_{t \rightarrow \infty} (-e^{-at} + 1) \\ &= 1\end{aligned}$$

$\therefore f(x)$ is a well defined PDF.

$$\begin{aligned}\text{c i} \quad \text{With } u' &= e^{-ax} & v &= ax \\ u &= \frac{1}{-a} e^{-ax} & v' &= a \\ \therefore \int axe^{-ax} dx &= uv - \int uv' \\ &= -xe^{-ax} - \int -e^{-ax} dx \\ &= -xe^{-ax} + \frac{1}{-a} e^{-ax} + \text{constant} \\ &= -e^{-ax} \left(x + \frac{1}{a} \right) + \text{constant}\end{aligned}$$

$$\begin{aligned}\text{ii} \quad \text{With } u' &= e^{-ax} & v &= ax^2 \\ u &= \frac{1}{-a} e^{-ax} & v' &= 2ax \\ \therefore \int ax^2 e^{-ax} dx &= -x^2 e^{-ax} - \int -2xe^{-ax} \\ &= -x^2 e^{-ax} + 2 \left(\frac{1}{a} \right) \times -e^{-ax} \left(x + \frac{1}{a} \right) + \text{constant} \\ &= e^{-ax} \left(-x^2 - \frac{2x}{a} - \frac{2}{a^2} \right) + \text{constant}\end{aligned}$$

$$\begin{aligned}\text{d} \quad \mu = E(X) &= \int_0^\infty xae^{-ax} dx \\ &= \left[-e^{-ax} \left(x + \frac{1}{a} \right) \right]_0^\infty \\ &= 0 - \left(-\frac{1}{a} \right) \\ &= \frac{1}{a} \\ E(X^2) &= \int_0^\infty x^2 ae^{-ax} dx \\ &= \left[e^{-ax} \left(-x^2 - \frac{2x}{a} - \frac{2}{a^2} \right) \right]_0^\infty \\ &= 0 - \left(-\frac{2}{a^2} \right) \\ &= \frac{2}{a^2}\end{aligned}$$

$$\text{and } \text{Var}(X) = E(X^2) - [E(X)]^2$$

$$\begin{aligned}&= \frac{2}{a^2} - \left(\frac{1}{a} \right)^2 \\ &= \frac{1}{a^2}\end{aligned}$$

EXERCISE D.1

1 a X has probability distribution:

x	1	2	3
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$\therefore X \sim \text{DU}\left(\frac{1}{3}\right)$$

$$\begin{aligned}\text{and } G(t) &= p_1 t + p_2 t^2 + p_3 t^3 \\ &= \frac{1}{3}(t + t^2 + t^3) \\ &= \frac{t}{3}(1 + t + t^2) \quad \text{or} \quad \frac{t(t^3 - 1)}{3(t - 1)} \quad \text{for } t \in \mathbb{R}\end{aligned}$$

- b** X has probability distribution:

x	1	2	5
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$\begin{aligned}\therefore G(t) &= \frac{1}{3}t + \frac{1}{3}t^2 + \frac{1}{3}t^5 \\ &= \frac{t}{3}(1 + t + t^4), \quad t \in \mathbb{R}\end{aligned}$$

c $G(t) = \frac{2}{11}t + \frac{3}{11}t^2 + \frac{5}{11}t^7 + \frac{1}{11}t^{12}$
 $= \frac{t}{11}(2 + 3t + 5t^6 + t^{11}), \quad t \in \mathbb{R}$

- 2 a** X has probability distribution:

x	1	2	3
$P(X = x)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$

b $G(t) = \frac{1}{5}t + \frac{2}{5}t^2 + \frac{2}{5}t^3$
 $= \frac{t}{5}(1 + 2t + 2t^2)$

- 3 a** $X \sim B(3, \frac{2}{5})$

X has probability distribution:

x	0	1	2	3
$P(X = x)$	$\frac{27}{125}$	$\frac{54}{125}$	$\frac{36}{125}$	$\frac{8}{125}$

b $G(t) = \frac{27}{125} + \frac{54}{125}t + \frac{36}{125}t^2 + \frac{8}{125}t^3$
 $= \frac{1}{125}(8t^3 + 36t^2 + 54t + 27)$
 $= \frac{1}{125}(2t + 3)^3$

- 4 a** $X \sim B(1, p)$ has probability distribution:

x	0	1
$P(X = x)$	$1 - p$	p

$$\begin{aligned}\therefore G(t) &= p_0 + p_1t \\ &= (1 - p) + pt \\ &= 1 - p + pt\end{aligned}$$

b For $X \sim B(1, 0.4)$, $G(t) = 0.6 + 0.4t$
 $= \frac{3 + 2t}{5}$

- 5 a** X has probability distribution:

x	0	1
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{2}$

$$\begin{aligned}\therefore G(t) &= \frac{1}{2}t^0 + \frac{1}{2}t^1 \\ &= \frac{1}{2}(1 + t)\end{aligned}$$

- b** Y has probability distribution:

y	0	1	2
$P(Y = y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$\begin{aligned}\therefore H(t) &= \frac{1}{4}t^0 + \frac{1}{2}t^1 + \frac{1}{4}t^2 \\ &= \frac{1}{4}(1 + 2t + t^2)\end{aligned}$$

c $H(t) = \frac{1}{4}(1 + t)^2 = [G(t)]^2$

- 6 a** $X \sim B(4, \frac{1}{6})$ and X has probability distribution

$$\begin{aligned}& \left(\frac{5}{6} + \frac{1}{6}\right)^4 \\ &= \left(\frac{5}{6}\right)^4 + 4\left(\frac{5}{6}\right)^3\left(\frac{1}{6}\right) + 6\left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right)^2 + 4\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^3 + \left(\frac{1}{6}\right)^4\end{aligned}$$

which is

x	0	1	2	3	4
$P(X = x)$	$\frac{625}{1296}$	$\frac{500}{1296}$	$\frac{150}{1296}$	$\frac{20}{1296}$	$\frac{1}{1296}$

$$\begin{aligned}\therefore H(t) &= \frac{625}{1296}t^0 + \frac{500}{1296}t^1 + \frac{150}{1296}t^2 + \frac{20}{1296}t^3 + \frac{1}{1296}t^4 \\ \therefore H(t) &= \frac{1}{1296}[625 + 500t + 150t^2 + 20t^3 + t^4]\end{aligned}$$

b $[G(t)]^4 = \left(\frac{5}{6} + \frac{1}{6}t\right)^4$
 $= \frac{1}{6^4}(5 + t)^4$
 $= \frac{1}{1296}(5^4 + 4(5)^3t + 6(5)^2t^2 + 4(5)t^3 + t^4)$
 $= \frac{1}{1296}(625 + 500t + 150t^2 + 20t^3 + t^4)$
 $= H(t)$

- 7 a** X has probability distribution

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\begin{aligned}\therefore G(t) &= \frac{1}{6}t + \frac{1}{6}t^2 + \frac{1}{6}t^3 + \frac{1}{6}t^4 + \frac{1}{6}t^5 + \frac{1}{6}t^6 \\ &= \frac{t}{6}(1 + t + t^2 + t^3 + t^4 + t^5)\end{aligned}$$

- b** Similarly for Y ,

$$H(t) = \frac{t}{4}(1 + t + t^2 + t^3)$$

- c i**

	4	5	6	7	8	9	10
Die 2	3	4	5	6	7	8	9
	2	3	4	5	6	7	8
	1	2	3	4	5	6	7
	1	2	3	4	5	6	

Die 1

$U = X + Y$ has probability distribution:

U	2	3	4	5	6	7	8	9	10
$P(U = u)$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{3}{24}$	$\frac{4}{24}$	$\frac{4}{24}$	$\frac{4}{24}$	$\frac{3}{24}$	$\frac{2}{24}$	$\frac{1}{24}$

ii $K(t) = \frac{1}{24}t^2 + \frac{2}{24}t^3 + \frac{3}{24}t^4 + \frac{4}{24}t^5 + \frac{4}{24}t^6 + \frac{4}{24}t^7$
 $+ \frac{3}{24}t^8 + \frac{2}{24}t^9 + \frac{1}{24}t^{10}$
 $= \frac{t^2}{24}[1 + 2t + 3t^2 + 4t^3 + 4t^4 + 4t^5 + 3t^6$
 $+ 2t^7 + t^8]$

- d** $G(t)H(t)$

$$\begin{aligned}&= \frac{t}{6}(1 + t + t^2 + t^3) \times \frac{t}{4}(1 + t + t^2 + t^3 + t^4 + t^5) \\ &= \frac{t^2}{24}(1 + t + t^2 + t^3)(1 + t + t^2 + t^3 + t^4 + t^5) \\ &= \frac{t^2}{24}[1 + 2t + 3t^2 + 4t^3 + 4t^4 + 4t^5 + 3t^6 + 2t^7 + t^8] \\ &\quad \text{(using synthetic multiplication)} \\ &= K(t)\end{aligned}$$

- 8 a** X is the number of independent trials needed to get a successful outcome. Each trial has probability $\frac{1}{4}$ of being successful

$$\therefore X \sim \text{Geo}\left(\frac{1}{4}\right)$$

- b i** $P(X = 1) = p = p_1 = \frac{1}{4}$

ii $P(X = 2) = p(1 - p) = \frac{1}{4} \times \frac{3}{4}$

iii $P(X = k) = \frac{1}{4}\left(\frac{3}{4}\right)^{k-1}$

$$\begin{aligned}
 \text{c } G(t) &= \sum_{k=1}^{\infty} p_k t^k \\
 &= \frac{1}{4}t + \frac{1}{4}\left(\frac{3}{4}\right)t^2 + \left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^2 t^3 + \dots \\
 &= \frac{1}{4}t \left(1 + \left(\frac{3}{4}\right)t + \left(\frac{3}{4}\right)^2 t^2 + \dots\right) \\
 &= \frac{1}{4}t \left(\frac{1}{1 - \frac{3}{4}t}\right) \quad \text{for } \left|\frac{3}{4}t\right| < 1 \\
 &= \frac{t}{4 - 3t} \quad \text{for } |t| < \frac{4}{3} \\
 \text{and the domain is } t \in]-\frac{4}{3}, \frac{4}{3}[
 \end{aligned}$$

EXERCISE D.2

- 1 $X \sim B(1, p)$ has probability distribution:

x	0	1
$P(X = x)$	$1 - p$	p

$$\begin{aligned}
 \therefore G(t) &= (1 - p)t^0 + pt^1 \\
 &= 1 - p + pt
 \end{aligned}$$

- 2 $X \sim \text{Po}(m)$ has $P(X = k) = \frac{m^k e^{-m}}{k!}$

$$\begin{aligned}
 \therefore G(t) &= \sum_{k=0}^{\infty} P(X = k)t^k \\
 &= \sum_{k=0}^{\infty} \frac{m^k e^{-m}}{k!} t^k \\
 &= \sum_{k=0}^{\infty} \frac{(mt)^k e^{-m}}{k!} \\
 &= e^{-m} \sum_{k=0}^{\infty} \frac{(mt)^k}{k!} \\
 &= e^{-m} \times e^{mt} \\
 &= e^{mt-m} \\
 &= e^{m(t-1)}, \quad t \in \mathbb{R}
 \end{aligned}$$

- 3 $X \sim B(n, p)$ has $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$

$$\begin{aligned}
 \therefore G(t) &= \sum_{x=0}^n P(X = x)t^x \\
 &= \sum_{x=0}^n \binom{n}{x} p^x (1 - p)^{n-x} t^x \\
 &= \sum_{x=0}^n \binom{n}{x} (pt)^x (1 - p)^{n-x} \\
 &= (pt + (1 - p))^n \\
 &= (1 - p + pt)^n, \quad n \in \mathbb{Z}^+
 \end{aligned}$$

- 4 a $X \sim \text{Po}(6)$ has $G(t) = e^{6(t-1)}$, $t \in \mathbb{R}$

b $X \sim B(10, 0.35)$ has $G(t) = (1 - 0.35 + 0.35t)^n$
 $= (0.65 + 0.35t)^n$, $n \in \mathbb{Z}^+$

c $X \sim \text{Geo}(0.7)$ has $G(t) = \frac{0.7t}{1 - 0.3t}$, $|t| < \frac{1}{0.3}$
 $= \frac{7t}{10 - 3t}$, $|t| < \frac{10}{3}$

d $X \sim \text{NB}(6, 0.2)$ has $G(t) = \left(\frac{0.2t}{1 - 0.8t}\right)^6$, $|t| < \frac{1}{0.8}$
 $= \left(\frac{t}{5 - 4t}\right)^6$, $|t| < \frac{5}{4}$

- 5 a If $X \sim B(n, p)$, $\mu = np$
and $G(t) = (1 - p + pt)^n$
 $= (1 + p(t - 1))^n$
 $= \left(1 + \frac{\mu(t - 1)}{n}\right)^n$

- b As n gets large $G(t) \rightarrow e^{\mu(t-1)}$

$$\left\{ \lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right) = e^a, \quad a \in \mathbb{R} \right\}$$

Since $e^{\mu(t-1)}$ is the Poisson PGF with $m = \mu$, for large n the Poisson probability distribution can be used to approximate the binomial probability distribution.

EXERCISE D.3

1 $G(t) = \frac{1}{2} + \frac{1}{2}t$
 $\therefore G'(t) = \frac{1}{2}$ and $G''(t) = 0$

$$E(X) = G'(1) = \frac{1}{2} \quad \text{and}$$

$$\begin{aligned}
 \text{Var}(X) &= G''(1) + G'(1) - [G'(1)]^2 \\
 &= 0 + \frac{1}{2} - \left(\frac{1}{2}\right)^2 \\
 &= \frac{1}{4}
 \end{aligned}$$

- 2 a X has probability distribution:

x	1	2	3
$P(X = x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

$$\therefore G(t) = \frac{1}{6}t + \frac{2}{6}t^2 + \frac{3}{6}t^3, \quad t \in \mathbb{R}$$

b $G'(t) = \frac{1}{6} + \frac{2}{3}t + \frac{3}{2}t^2$ and
 $G''(t) = \frac{2}{3} + 3t$

$$\text{Thus } E(X) = G'(1) = \frac{1}{6} + \frac{2}{3} + \frac{3}{2}$$

$$\therefore E(X) = 2\frac{1}{3}$$

$$\begin{aligned}
 \text{and } \text{Var}(X) &= G''(1) + G'(1) - [G'(1)]^2 \\
 &= 3\frac{2}{3} + 2\frac{1}{3} - \left(2\frac{1}{3}\right)^2 \\
 &= 11\frac{4}{9}
 \end{aligned}$$

- 3 a $X \sim \text{Po}(12)$, $\therefore G(t) = e^{12(t-1)}$

b $G'(t) = 12e^{12(t-1)}$ and $G''(t) = 144e^{12(t-1)}$

$$\therefore G'(1) = 12 \quad \text{and} \quad G''(1) = 144$$

$$\begin{aligned}
 \therefore E(X) &= 12 \quad \text{and} \quad \text{Var}(X) = 144 + 12 - 12^2 \\
 &= 12
 \end{aligned}$$

- 4 $X \sim B(n, p)$

$$\therefore G(t) = (1 - p + pt)^n$$

$$\therefore G'(t) = n(1 - p + pt)^{n-1} \times p$$

$$\begin{aligned}
 \therefore E(X) &= G'(1) \\
 &= n(1)^{n-1}p \\
 &= np
 \end{aligned}$$

$$\text{and } G''(t) = np(n-1)(1 - p + pt)^{n-2} \times p$$

$$\begin{aligned}
 \therefore G''(1) &= n(n-1)p^2 \times (1) \\
 &= n(n-1)p^2
 \end{aligned}$$

$$\text{Now } \text{Var}(X)$$

$$\begin{aligned}
 &= G''(1) + G'(1) - [G'(1)]^2 \\
 &= n(n-1)p^2 + np - n^2p^2 \\
 &= np[np - p + 1 - np] \\
 &= np(1 - p)
 \end{aligned}$$

5 $X \sim \text{Geo}(p)$

$$\therefore G(t) = \frac{pt}{1 - t(1-p)}$$

$$\begin{aligned} G'(t) &= \frac{p[1 - t(1-p)] - pt[p-1]}{[1 - t(1-p)]^2} \\ &= \frac{p - \cancel{pt} + \cancel{p^2t} - \cancel{p^2t} + \cancel{pt}}{[1 - t(1-p)]^2} \\ &= \frac{p}{[1 - t(1-p)]^2} \end{aligned}$$

Since $G'(t) = p[1 - t(1-p)]^{-2}$

$$\begin{aligned} G''(t) &= -2p[1 - t(1-p)]^{-3}(p-1) \\ &= \frac{-2p(p-1)}{[1 - t(1-p)]^3} \end{aligned}$$

Thus $G'(1) = \frac{p}{p^2} = \frac{1}{p}$ and $G''(1) = \frac{-2p(p-1)}{p^3}$

$$\therefore E(X) = \frac{1}{p}$$

and $\text{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$

$$\begin{aligned} &= \frac{-2p(p-1)}{p^3} + \frac{1}{p} - \frac{1}{p^2} \\ &= -\frac{2}{p} + \frac{2}{p^2} + \frac{1}{p} - \frac{1}{p^2} \\ &= \frac{1}{p^2} - \frac{1}{p} \\ &= \frac{1-p}{p^2} \end{aligned}$$

6 $X \sim \text{Po}(m)$

$$\therefore G(t) = e^{m(t-1)}$$

Now $G'(t) = me^{m(t-1)}$

and $G''(t) = m^2e^{m(t-1)}$

$$\therefore E(X) = G'(1) = me^0 = m$$

$$\begin{aligned} \text{and } \text{Var}(X) &= G''(1) + G'(1) - [G'(1)]^2 \\ &= m^2 + m - m^2 \\ &= m \end{aligned}$$

7 $X \sim \text{NB}(r, p)$

$$\therefore G(t) = \left(\frac{pt}{1 - t(1-p)} \right)^r$$

$$\begin{aligned} \therefore G'(t) &= r \left(\frac{pt}{1 - t(1-p)} \right)^{r-1} \left[\frac{p[1 - t(1-p)] - pt(p-1)}{[1 - t(1-p)]^2} \right] \\ &= r \left(\frac{pt}{1 - t(1-p)} \right)^{r-1} \left[\frac{p - \cancel{pt} + \cancel{p^2t} - \cancel{p^2t} + \cancel{pt}}{[1 - t(1-p)]^2} \right] \\ &= r \left(\frac{pt}{1 - t(1-p)} \right)^{r-1} \times \frac{p}{[1 - t(1-p)]^2} \end{aligned}$$

$$\begin{aligned} \therefore G'(1) &= r \left(\frac{p}{p} \right)^{r-1} \times \frac{p}{p^2} \\ \therefore E(X) &= \frac{r}{p} \end{aligned}$$

$$\begin{aligned} G''(t) &= r(r-1) \left(\frac{pt}{1 - t(1-p)} \right)^{r-2} \times \frac{p^2}{[1 - t(1-p)]^4} \\ &\quad + r \left(\frac{pt}{1 - t(1-p)} \right)^{r-1} \times -2p[1 - t(1-p)]^{-3}(p-1) \end{aligned}$$

$$\begin{aligned} \therefore G''(1) &= \frac{r(r-1)p^2}{p^4} - \frac{2p(p-1)r}{p^3} \\ &= \frac{r(r-1)}{p^2} - \frac{2r(p-1)}{p^2} \\ &= \frac{r^2 + r - 2pr}{p^2} \end{aligned}$$

$$\begin{aligned} \therefore \text{Var}(X) &= G''(1) + G'(1) - [G'(1)]^2 \\ &= \frac{r^2 + r - 2pr}{p^2} + \frac{r}{p} - \frac{r^2}{p^2} \\ &= \frac{\cancel{r^2} + r - 2pr + rp - \cancel{r^2}}{p^2} \\ &= \frac{r(1-p)}{p^2} \end{aligned}$$

Condition: $|t| < \frac{1}{1-p}$

EXERCISE D.4

1 a X has probability distribution:

x	1	2
$P(X=x)$	$\frac{1}{2}$	$\frac{1}{2}$

$$\therefore G(t) = \frac{1}{2}t^1 + \frac{1}{2}t^2$$

$$\therefore G(t) = \frac{1}{2}t(1+t)$$

b Y has probability distribution:

y	1	2	3	4
$P(Y=y)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$$\begin{aligned} \therefore H(t) &= \frac{1}{4}t + \frac{1}{4}t^2 + \frac{1}{4}t^3 + \frac{1}{4}t^4 \\ &= \frac{t}{4}(1+t+t^2+t^3) \end{aligned}$$

c i U has PGF $G(t)H(t)$

$$\begin{aligned} &= \frac{t^2}{8}(1+t)(1+t+t^2+t^3) \\ &= \frac{t^2}{8}(1+2t+2t^2+2t^3+t^4) \\ &= \frac{t^2}{8} + \frac{t^3}{4} + \frac{t^4}{4} + \frac{t^5}{4} + \frac{t^6}{8} \end{aligned}$$

ii $P(U=4)$ = the coefficient of t^4
 $= \frac{1}{4}$

2 a Let R = result of rolling the tetrahedral die

r	1	2	3	4
$P(R=r)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$$\begin{aligned} \therefore G(t) &= \frac{1}{4}t + \frac{1}{4}t^2 + \frac{1}{4}t^3 + \frac{1}{4}t^4 \\ &= \frac{t}{4}(1+t+t^2+t^3) \end{aligned}$$

and if S = result of rolling the 6-sided die

$$H(t) = \frac{t}{6}(1+t+t^2+t^3+t^4+t^5)$$

Thus if $X = R + S$, the PGF for X is

$$\begin{aligned} & \frac{t}{4}(1+t+t^2+t^3) \times \frac{t}{6}(1+t+t^2+t^3+t^4+t^5) \\ &= \frac{t^2}{24}(1+2t+3t^2+4t^3+4t^4+4t^5+3t^6+2t^7+t^8) \end{aligned}$$

b $P(X=5)$ = coefficient of t^5

$$\begin{aligned} &= \frac{1}{24} \times 4 \\ &= \frac{1}{6} \end{aligned}$$

3 $X \sim \text{Po}(m)$ and $Y \sim \text{Po}(\lambda)$

a Consider $U = X + Y$
 U has PGF $G_X(t)G_Y(t)$
 $= e^{m(t-1)}e^{\lambda(t-1)}$
 $= e^{m(t-1)+\lambda(t-1)}$
 $= e^{[m+\lambda](t-1)}$

b This is the PGF for a Poisson probability function with parameter $m + \lambda$
 So, $X + Y \sim \text{Po}(m + \lambda)$.

4 $X \sim B(n, p)$, $Y \sim B(m, p)$

a $U = X + Y$ has PGF
 $G_X(t)G_Y(t)$
 $= (1-p+pt)^n(1-p+pt)^m$
 $= (1-p+pt)^{n+m}$

b This is the PGF for a Binomial probability function with parameters p and $n + m$
 So, $X + Y \sim B(n + m, p)$.

5 $X \sim \text{Geo}(p)$, $Y \sim \text{Geo}(p)$

a $U = X + Y$ has PGF
 $G_X(t)G_Y(t) = \frac{pt}{1-t(1-p)} \times \frac{pt}{1-t(1-p)}$
 $= \frac{p^2t^2}{[1-t(1-p)]^2}$
 $= \left(\frac{pt}{1-t(1-p)} \right)^2$

b This is the PGF for a Negative binomial probability function with parameters $r = 2$ and p
 So, $X + Y \sim \text{NB}(2, p)$.

6 $X \sim \text{NB}(r, p)$ and $Y \sim \text{NB}(s, p)$

a $U = X + Y$ has PGF
 $G_X(t)G_Y(t) = \left(\frac{pt}{1-t(1-p)} \right)^r \left(\frac{pt}{1-t(1-p)} \right)^s$
 $= \left(\frac{pt}{1-t(1-p)} \right)^{r+s}$

b This is the PGF for a Negative binomial probability distribution with parameters $r + s$ and p
 So, $X + Y \sim \text{NB}(r + s, p)$.

7 a $P(a=4) = \frac{1}{7}$

b $Y \sim \text{Geo}(\frac{1}{7})$

i PGF of Y is $G(t) = \frac{\frac{1}{7}t}{1-t(\frac{6}{7})}$ for $|t| < \frac{1}{6}$

$$\therefore G(t) = \frac{t}{7-6t} \text{ for } |t| < \frac{1}{6}$$

$$\begin{aligned} \text{ii } G'(t) &= \frac{1(7-6t) - t(-6)}{(7-6t)^2} \\ &= \frac{7}{(7-6t)^2} \end{aligned}$$

$$\therefore E(Y) = G'(1) = \frac{7}{1^2} = 7$$

$$\begin{aligned} \text{Since } G'(t) &= 7(7-6t)^{-2} \\ G''(t) &= -14(7-6t)^{-3}(-6) \\ &= \frac{84}{(7-6t)^3} \end{aligned}$$

$$\begin{aligned} \text{Var}(Y) &= G''(1) + G'(1) - [G'(1)]^2 \\ &= 84 + 7 - 7^2 \\ &= 42 \end{aligned}$$

c $X = \text{NB}(3, \frac{1}{7})$

i Y_1 = number of spins needed to get the first '4'
 Y_2 = number of spins needed to get the second '4'
 Y_3 = number of spins needed to get the third '4'
 $\therefore$ each Y_i are geometric with parameter $\frac{1}{7}$.

$$\begin{aligned} \text{ii } E(X) &= E(Y_1) + E(Y_2) + E(Y_3) \\ &= 7 + 7 + 7 \\ &= 21 \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= \text{Var}(Y_1) + \text{Var}(Y_2) + \text{Var}(Y_3) \\ &= 42 + 42 + 42 \\ &= 126 \end{aligned}$$

iii If Y_1 , Y_2 , and Y_3 are discrete random variables with values in $\mathbb{Z}^+$ and PGFs $G_{Y_1}(t)$, $G_{Y_2}(t)$, and $G_{Y_3}(t)$ then $Y_1 + Y_2 + Y_3$ has PGF $G_{Y_1}(t)G_{Y_2}(t)G_{Y_3}(t)$ with mean $E(Y_1) + E(Y_2) + E(Y_3)$ and $\text{Var}(Y_1) + \text{Var}(Y_2) + \text{Var}(Y_3)$.

8 a i $G(t) = 1 - p + pt$
 $\therefore G'(t) = p$
 and $G'(1) = p$
 $\therefore E(Y) = p$

ii Since $G'(t) = p$, $G''(t) = 0$
 $\therefore \text{Var}(Y) = G''(1) + G'(1) - [G'(1)]^2$
 $= 0 + p - p^2$
 $= p(1-p)$

b i $X \sim B(5, \frac{1}{6})$ ($p = \frac{1}{6}$)
 If $X = Y_1 + Y_2 + Y_3 + Y_4 + Y_5$ where each $Y_i \sim B(1, \frac{1}{6})$ then

$$\begin{aligned} H(t) &= G(Y_1)G(Y_2)G(Y_3)G(Y_4)G(Y_5) \\ &= [G(Y)]^5 \\ &= (1-p+pt)^5 \end{aligned}$$

$$\therefore H(t) = (\frac{5}{6} + \frac{1}{6}t)^5$$

$$\begin{aligned} \text{ii } H'(t) &= 5(\frac{5}{6} + \frac{1}{6}t)^4 \times \frac{1}{6} \\ H''(t) &= \frac{5}{6} \times 4(\frac{5}{6} + \frac{1}{6}t)^3 \times \frac{1}{6} \\ \therefore E(X) &= H'(1) = \frac{5}{6} \end{aligned}$$

$$\begin{aligned} \text{and } \text{Var}(X) &= H''(1) + H'(1) - [H'(1)]^2 \\ &= \frac{20}{36} + \frac{5}{6} - \frac{25}{36} \\ &= \frac{25}{36} \end{aligned}$$

$$\begin{aligned}
 \text{iii } X &= Y_1 + Y_2 + Y_3 + Y_4 + Y_5 \text{ where } Y_i \sim B(1, \frac{1}{6}) \\
 \therefore E(X) &= E(Y_1) + E(Y_2) + \dots + E(Y_5) \\
 &= 5p \\
 &= 5 \times \frac{1}{6} \\
 &= \frac{5}{6}
 \end{aligned}$$

$$\begin{aligned}
 \text{and } \text{Var}(Y) &= \sum_{i=1}^5 \text{Var}(Y_i) \\
 &= \sum_{i=1}^5 (\frac{1}{6} \times \frac{5}{6}) \\
 &= 5 \times \frac{5}{36} \\
 &= \frac{25}{36}
 \end{aligned}$$

9 $Y_i \sim \text{Geo}(p)$ for $i = 1, 2, 3, \dots, r$

$X = Y_1 + Y_2 + Y_3 + \dots + Y_r$ is the sum of r independent random variables

$$\begin{aligned}
 \therefore G(X) &= G(Y_1)G(Y_2)G(Y_3)\dots G(Y_r) \\
 &= \left(\frac{pt}{1-t(1-p)} \right)^r, \quad |t| < \frac{1}{1-p}
 \end{aligned}$$

which is the Negative binomial PGF as expected.

10 X and Y are independent discrete random variables.

X has PGF $G_X(t)$ and Y has PGF $G_Y(t)$.

$\therefore X + Y$ has PGF $G_X(t)G_Y(t) = G(t)$.

$$\begin{aligned}
 \text{a } G'(t) &= G'_X(t)G_Y(t) + G_X(t)G'_Y(t) \\
 \therefore G'(1) &= G'_X(1)G_Y(1) + G_X(1)G'_Y(1) \\
 &= E(X) \times 1 + 1 \times E(Y) \\
 &= E(X) + E(Y)
 \end{aligned}$$

Thus $E(X + Y) = E(X) + E(Y)$

$$\begin{aligned}
 \text{b Also } G''(t) &= G''_X(t)G_Y(t) + G'_X(t)G'_Y(t) \\
 &\quad + G'_X(t)G'_Y(t) + G_X(t)G''_Y(t) \\
 \therefore G''(1) &= G''_X(1)G_Y(1) + 2G'_X(1)G'_Y(1) \\
 &\quad + G_X(1)G''_Y(1) \\
 &= G''_X(1) + 2G'_X(1)G'_Y(1) + G''_Y(1)
 \end{aligned}$$

Now $\text{Var}(X + Y)$

$$\begin{aligned}
 &= G''(1) + G'(1) - [G'(1)]^2 \\
 &= G''_X(1) + 2G'_X(1)G'_Y(1) + G''_Y(1) \\
 &\quad + G'_X(1) + G'_Y(1) - [G'_X(1) + G'_Y(1)]^2 \\
 &= G''_X(1) + \cancel{2G'_X(1)G'_Y(1)} + G''_Y(1) + G'_X(1) \\
 &\quad + G'_Y(1) - [G'_X(1)]^2 - \cancel{2G'_X(1)G'_Y(1)} \\
 &\quad - [G'_Y(1)]^2 \\
 &= G''_X(1) + G'_X(1) - [G'_X(1)]^2 + G''_Y(1) + G'_Y(1) \\
 &\quad - [G'_Y(1)]^2 \\
 &= \text{Var}(X) + \text{Var}(Y)
 \end{aligned}$$

11 a X has PGF $G(t) = p_0 + p_1t + p_2t^2 + p_3t^3 + \dots$ for X values of 0, 1, 2, 3, 4,

i $X + 2$ takes values 2, 3, 4, 5, 6,

$$\begin{aligned}
 H(t) &= p_0t^2 + p_1t^3 + p_2t^4 + p_3t^5 + \dots \\
 &= t^2(p_0 + p_1t + p_2t^2 + p_3t^3 + \dots) \\
 &= t^2G(t)
 \end{aligned}$$

ii $3X$ takes values 0, 3, 6, 9, 12,

$$\begin{aligned}
 H(t) &= p_0 + p_1t^3 + p_2t^6 + p_3t^9 + \dots \\
 &= G(t^3)
 \end{aligned}$$

b i $aX + b$ takes values $b, a + b, 2a + b, 3a + b, \dots$

$$\begin{aligned}
 \therefore H(t) &= p_0t^b + p_1t^{a+b} + p_2t^{2a+b} + \dots \\
 &= t^b[p_0 + p_1t^a + p_2t^{2a} + \dots] \\
 &= t^bG(t^a)
 \end{aligned}$$

ii $H(t) = t^bG(t^a)$

$$\begin{aligned}
 \therefore H'(t) &= bt^{b-1}G(t^a) + t^bG'(t^a)at^{a-1} \\
 &= bt^{b-1}G(t^a) + at^{a+b-1}G'(t^a)
 \end{aligned}$$

Now $E(aX + b) = H'(1) = bG(1) + aG'(1)$

$$\begin{aligned}
 \therefore E(aX + b) &= b \times 1 + aE(X) \\
 &= aE(X) + b
 \end{aligned}$$

Also,

$$\begin{aligned}
 H''(t) &= b(b-1)t^{b-2}G(t^a) + bt^{b-1}G'(t^a)at^{a-1} \\
 &\quad + a(a+b-1)t^{a+b-2}G'(t^a) \\
 &\quad + at^{a+b-1}G''(t^a)at^{a-1}
 \end{aligned}$$

$$\begin{aligned}
 \therefore H''(1) &= b(b-1)G(1) + abG'(1) \\
 &\quad + a(a+b-1)G'(1) + a^2G''(1) \\
 &= a^2G''(1) + (a^2 + 2ab - a)G'(1) + (b^2 - b)
 \end{aligned}$$

Now $\text{Var}(aX + b)$

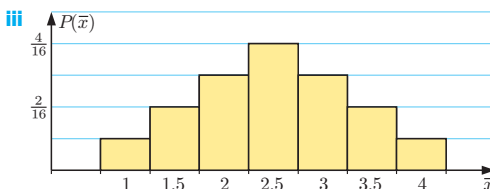
$$\begin{aligned}
 &= H''(1) + H'(1) - [H'(1)]^2 \\
 &= a^2G''(1) + (a^2 + 2ab - a)G'(1) + b^2 - b \\
 &\quad + b + aG'(1) - [b^2 + 2abG'(1) + [G'(1)]^2] \\
 &= a^2G''(1) + (a^2 + 2a\cancel{b} - \cancel{a} + \cancel{a} - 2a\cancel{b})G'(1) \\
 &\quad + \cancel{b^2} - \cancel{b} + \cancel{b} - \cancel{b^2} - a^2[G'(1)]^2 \\
 &= a^2(G''(1) + G'(1) - [G'(1)]^2) \\
 &= a^2\text{Var}(X)
 \end{aligned}$$

EXERCISE E.1

1 a i

Poss. sample	$\bar{x}$	Poss. sample	$\bar{x}$
1, 1	1	3, 1	2
1, 2	1.5	3, 2	2.5
1, 3	2	3, 3	3
1, 4	2.5	3, 4	3.5
2, 1	1.5	4, 1	2.5
2, 2	2	4, 2	3
2, 3	2.5	4, 3	3.5
2, 4	3	4, 4	4

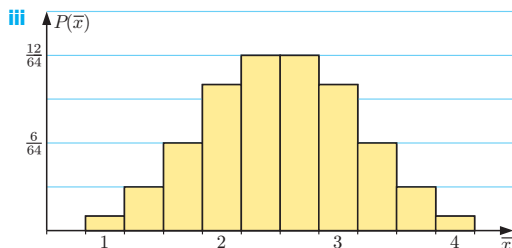
$\bar{x}$	1	1.5	2	2.5	3	3.5	4
Freq.	1	2	3	4	3	2	1
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$



b ii

$\bar{x}$	1	$\frac{4}{3}$	$\frac{5}{3}$	2	$\frac{7}{3}$
Freq.	1	3	6	10	12
$P(\bar{x})$	$\frac{1}{64}$	$\frac{3}{64}$	$\frac{6}{64}$	$\frac{10}{64}$	$\frac{12}{64}$

$\bar{x}$	$\frac{8}{3}$	3	$\frac{10}{3}$	$\frac{11}{3}$	4
Freq.	12	10	6	3	1
$P(\bar{x})$	$\frac{12}{64}$	$\frac{10}{64}$	$\frac{6}{64}$	$\frac{3}{64}$	$\frac{1}{64}$



Poss. sample	$\bar{x}$	Poss. sample	$\bar{x}$
2, 2, 2, 2	2	3, 3, 2, 2	$\frac{10}{4}$
2, 2, 2, 3	$\frac{9}{4}$	3, 2, 3, 2	$\frac{10}{4}$
2, 2, 3, 2	$\frac{9}{4}$	3, 2, 2, 3	$\frac{10}{4}$
2, 3, 2, 2	$\frac{9}{4}$	2, 3, 3, 3	$\frac{11}{4}$
3, 2, 2, 2	$\frac{9}{4}$	3, 2, 3, 3	$\frac{11}{4}$
2, 2, 3, 3	$\frac{10}{4}$	3, 3, 2, 3	$\frac{11}{4}$
2, 3, 2, 3	$\frac{10}{4}$	3, 3, 3, 2	$\frac{11}{4}$
2, 3, 3, 2	$\frac{10}{4}$	3, 3, 3, 3	3

$\bar{x}$	2	$\frac{9}{4}$	$\frac{10}{4}$	$\frac{11}{4}$	3
Freq.	1	4	6	4	1
$P(\bar{x})$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

$\bar{x}$	1	1.5	2	2.5	3	3.5
Freq.	1	2	3	4	5	6
$P(\bar{x})$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$

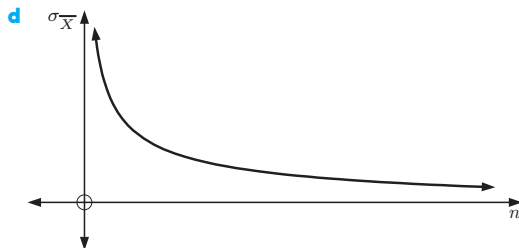
$\bar{x}$	4	4.5	5	5.5	6
Freq.	5	4	3	2	1
$P(\bar{x})$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

EXERCISE E.2

1 a $\mu_{\bar{X}} = \mu = 64$ b $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{36}} \approx 1.67$

2 a $\sigma_{\bar{X}} = \frac{24}{\sqrt{n}}$ b i 12 ii 6 iii 3

c If $\sigma_{\bar{X}} = 4$, then $\frac{24}{\sqrt{n}} = 4$
 $\therefore \sqrt{n} = 6$
 $\therefore n = 36$



e As n gets larger, $\sigma_{\bar{X}}$ gets smaller and approaches 0. Hence, for large n , $n \rightarrow$ population size, and the sampling error of the mean is effectively zero, i.e., when the sample is the population $\bar{x} = \mu$ without error.

3 $X \sim \text{Po}(6)$ $\therefore \mu = 6$ and $\sigma^2 = 6$

By the CLT, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ approximately

$$\therefore \bar{X} \sim N\left(6, \frac{6}{n}\right)$$

We require n such that $P(\bar{X} < 5) = 0.09$

$$\therefore P\left(\frac{\bar{X} - 6}{\sqrt{\frac{6}{n}}} < \frac{5 - 6}{\sqrt{\frac{6}{n}}}\right) < 0.09$$

$$\therefore P\left(Z < \frac{-1}{\sqrt{\frac{6}{n}}}\right) < 0.09$$

$$\therefore -\sqrt{\frac{n}{6}} = -1.3408$$

$$\therefore \frac{n}{6} = 1.798$$

$$\therefore n \approx 10.8$$

$$\therefore n \approx 11$$

4 $\mu = 100$ and $\sigma = 15$

a By the CLT, we expect $\mu_{\bar{X}} = 100$

b By the CLT, we expect $\sigma_{\bar{X}} = \frac{15}{\sqrt{36}} = 2.5$

c As n is sufficiently large we expect the distribution to be normal.

5 a $\mu = \sum p_i x_i$ $\sigma^2 = \sum p_i x_i^2 - \mu^2$
 $= 0 \times \frac{1}{2} + 1 \times \frac{1}{2}$ $= \frac{1}{2} \times 0 + \frac{1}{2} \times 1 - \frac{1}{4}$
 $= \frac{1}{2}$ $= \frac{1}{4}$
 $\therefore \sigma = \frac{1}{2}$

i	TTTT	TTTH	TTHH	HHHT	HHHH
		TTHT	THTH	HHTH	
		THTT	THHT	HTHH	
		HTTT	HHTT	THHH	
			HTHT		
			HTTH		

$\bar{X}_i$	0	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	1
p_i	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

ii $\mu_{\bar{X}} = \frac{1}{16}(0) + \frac{4}{16}\left(\frac{1}{4}\right) + \frac{6}{16}\left(\frac{2}{4}\right) + \frac{4}{16}\left(\frac{3}{4}\right) + \frac{1}{16}(1)$
 $= \frac{32}{64}$
 $= \frac{1}{2}$

$$\sigma_{\bar{X}}^2 = \frac{1}{16}(0)^2 + \frac{4}{16}\left(\frac{1}{4}\right)^2 + \frac{6}{16}\left(\frac{2}{4}\right)^2 + \frac{4}{16}\left(\frac{3}{4}\right)^2 + \frac{1}{16}(1)^2 - \left(\frac{1}{2}\right)^2$$

$$= \frac{1}{16}$$

Hence $\sigma_{\bar{X}} = \frac{1}{4}$

iii $\mu_{\bar{X}} = \mu = \frac{1}{2}$ ✓

$$\sigma_{\bar{X}} = \frac{1}{4} \text{ and } \frac{\sigma}{\sqrt{n}} = \frac{\frac{1}{2}}{\sqrt{4}} = \frac{1}{4} \quad \checkmark$$

6 a Let X be the value of a home in this suburb.
 $\mu = \text{€}620\,000$ with $\sigma = \text{€}80\,000$.

By the CLT, $\bar{X} \sim N\left(620\,000, \frac{80\,000^2}{25}\right)$ for large n .

$$P(\bar{X} \geq 643\,000) \approx 0.0753$$

- b** The answer may not be all that reliable as X is not normal. Hence, we treat the answer with great caution. Note that the result states that about 7.53% of all samples of size 25 will have an average value of at least €643 000.

- 7** Let H = heights of plants (in cm)

$$\mu_H = 21 \text{ and } \sigma_H = \sqrt{90} \text{ cm}$$

$$\bar{H} \sim N(21, \frac{90}{40}) \quad \{\text{CLT}\}$$

$$\text{as } \mu_{\bar{H}} = 21, \sigma_{\bar{H}}^2 = \frac{9}{4} \therefore \sigma_{\bar{H}} = \frac{3}{2}$$

$$P(19.5 < \bar{H} < 24) \approx 0.819$$

- 8** Let M = mass of male students (in kg).

$$\mu_M = 70 \text{ and } \sigma_M = 5$$

$$\bar{M} \sim N(70, \frac{5^2}{64}) \quad \{\text{CLT}\}$$

$$\text{as } \mu_{\bar{M}} = 70 \text{ and } \sigma_{\bar{M}}^2 = \frac{5^2}{64} = (\frac{5}{8})^2$$

$$\therefore P(\bar{M} < 68.75) \approx 0.0228$$

- 9** $X \sim B(20, 0.6)$

$$E(X) = np \quad \text{Var}(X) = np(1-p)$$

$$= 12 \quad = 4.8$$

$$\bar{X} \sim N(12, \frac{4.8}{100}) \quad \{\text{CLT}\}$$

$$\mu_{\bar{X}} = 12 \text{ and } \sigma_{\bar{X}} = \sqrt{\frac{4.8}{100}}$$

$$\text{a } P(\bar{X} > 12.4) \approx 0.0339$$

$$\text{b } P(\bar{X} < 12.2) \approx 0.819$$

- 10** Let X = duration of pregnancy (in days)

$$X \sim N(267, 15^2)$$

$$\text{a } P(\text{overdue between 7 and 14 days})$$

$$= P(274 < X < 281)$$

$$\approx 0.145$$

$$\therefore \text{about 14.5\% are overdue.}$$

$$\text{b } \text{We need to find } k \text{ such that } P(X \leq k) = 0.8$$

$$k \approx 279.6$$

So, the longest 20% of pregnancies last 280 days or more.

$$\text{c } \text{i } \bar{X} \sim N(267, \frac{15^2}{64})$$

$$\text{ii normal with mean 267 and } \sigma = \frac{15}{8} \text{ days.}$$

$$\text{iii } P(\bar{X} \leq 260) \approx 0.000\,094\,5 \text{ which is a very small chance.}$$

$$\text{d As } X \text{ is now not normally distributed the answer to a and b above are not acceptable.}$$

However as $n > 30$ the answers to **c ii** and **iii** are still good approximations.

- 11** Let A = units of milk from an Ayrshire cow.

J = units of milk from a Jersey cow.

$$A \sim N(49, 5.87^2) \text{ and } J \sim N(44.8, 5.12^2)$$

$$\text{a } P(A > 50) \approx 0.432$$

$$\text{b Consider } D = J - A$$

$$E(D) = E(J) - E(A) \quad \text{Var}(D) = \text{Var}(J) + \text{Var}(A)$$

$$= 44.8 - 49$$

$$= 5.87^2 + 5.12^2$$

$$= -4.2$$

$$\approx 60.67$$

Now if J and A are independent variables

$$D \sim N(-4.2, 60.27) \quad \{\sigma_D = \sqrt{60.27}\}$$

$$\therefore P(D > 0) \approx 0.295$$

$$\text{c } \bar{J} \sim N(44.8, \frac{5.12^2}{25}) \quad \{\sigma = \frac{5.12}{5}\}$$

$$\therefore P(\bar{J} > 46) \approx 0.121$$

$$\text{d } \bar{J} \sim N(44.8, \frac{5.12^2}{25}), \bar{A} \sim N(49, \frac{5.87^2}{15})$$

Consider $U = \bar{A} - \bar{J}$

$$E(U) = E(\bar{A}) - E(\bar{J}) \quad \sigma_U^2 = \text{Var}(\bar{A}) + \text{Var}(\bar{J})$$

$$= 49 - 44.8$$

$$= \frac{5.87^2}{15} + \frac{5.12^2}{25}$$

$$= 4.2$$

$$\approx 3.3457$$

Assuming $\bar{A}$ and $\bar{J}$ are independent

$$P(U > 4) \approx 0.544 \quad \{\sigma_U = \sqrt{3.3457}\}$$

- 12** Let W = weight of adult males

$$W \sim N(73.5, 8.24^2)$$

$$\text{If } n = 9, \bar{W} \sim N(73.5, \frac{8.24^2}{9}) \text{ and } P(\bar{W} \leq \frac{650}{9}) \approx 0.321$$

which is $\approx 32.1\%$

$$\text{If } n = 8, \bar{W} \sim N(73.5, \frac{8.24^2}{8})$$

$$\text{and } P(\bar{W} \leq \frac{650}{8}) \approx 0.996 \text{ or } 99.6\%$$

Thus the maximum recommended number is 8.

Note: We do not have to have n large here as W is already normally distributed.

- 13** $\mu_X = 74$ and $\sigma_X = 6$

$$\bar{X} \sim N\left(74, \frac{6^2}{n}\right)$$

$$P(\bar{X} < 70.4) = 0.001\,35$$

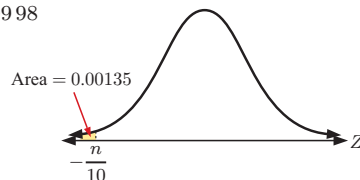
$$\therefore P\left(\frac{\bar{X} - 74}{\frac{6^2}{n}} < \frac{70.4 - 74}{\frac{6^2}{n}}\right) = 0.001\,35$$

$$\therefore P\left(Z < \frac{-3.6n}{6^2}\right) = 0.001\,35$$

$$\therefore P\left(Z < -\frac{n}{10}\right) = 0.001\,35$$

$$\therefore -\frac{n}{10} \approx -2.999\,98$$

$$\therefore n \approx 30$$



- 14** $\sigma_X = 4.55$ kg

Since $n = 100$ which is much > 30

$\therefore$ the CLT applies.

As the error could be positive or negative we need to find

$$P(|\bar{X} - \mu| < 0.8)$$

$$= P(-0.8 < \bar{X} - \mu < 0.8)$$

$$= P\left(\frac{-0.8}{\frac{\sigma}{\sqrt{100}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{100}}} < \frac{0.8}{\frac{\sigma}{\sqrt{100}}}\right)$$

$$= P\left(-\frac{8}{4.55} < Z < \frac{8}{4.55}\right)$$

$$\approx 0.921$$

- 15** $\mu_X = 100, \sigma_X = 15$

$$\bar{X} \sim N(100, \frac{15^2}{60}) \quad \{\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\}$$

$$P(\bar{X} \geq 105) \approx 0.004\,91$$

- 16 X = weight of a bar of chocolate

$$X \sim N(18.2, 3.3^2)$$

a $\mu_{\bar{X}} = 18.2$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3.3}{\sqrt{25}}$

- b i Claim is: each bar averages 17 g

ii Since $\bar{X} \sim N(18.2, \frac{3.3^2}{25})$,

$$P(\bar{X} < 17) \approx 0.0345$$

$\therefore$ 3.45% fail to meet the claim.

c i $\mu_{\bar{X}} = 18.2$ and $\sigma_{\bar{X}} = \frac{3.3}{\sqrt{26}}$

ii Since $\bar{X} \sim N(18.2, \frac{3.3^2}{26})$,

$$P(\bar{X} < \frac{425}{26}) \approx 0.00209$$

$\therefore$ about 0.209% fail to meet the claim now.

17 a i $\bar{S}$ has $\mu_{\bar{S}} = 315$ and $\sigma_{\bar{S}} = \frac{2}{\sqrt{15}}$

By the CLT $\bar{S} \sim N(315, \frac{4}{15})$

ii $\bar{E}$ has $\mu_{\bar{E}} = 950$ and $\sigma_{\bar{E}} = \frac{5}{\sqrt{10}}$

$$\therefore \bar{E} \sim N(950, \frac{25}{10})$$

b $P(\bar{E} < 3\bar{S})$

$$= P(\bar{E} - 3\bar{S} < 0)$$

$$\text{Let } U = \bar{E} - 3\bar{S}$$

$$\therefore E(U) = E(\bar{E}) - 3E(\bar{S})$$

$$= 950 - 3(315)$$

$$= 5$$

$$\text{and } \text{Var}(U) = \text{Var}(\bar{E}) + 9\text{Var}(\bar{S})$$

$$= \frac{25}{10} + 9(\frac{4}{15})$$

$$= 4.9$$

$$U \sim N(5, 4.9) \text{ and } P(U < 0) \approx 0.0119$$

- c Let the contents of the three small cartons be S_1 , S_2 , and S_3 and consider

$$V = E - (S_1 + S_2 + S_3)$$

$$E(V) = E(E) - E(S_1) - E(S_2) - E(S_3)$$

$$= 950 - 315 - 315 - 315$$

$$= 5$$

$$\text{Var}(V) = \text{Var}(E) + \text{Var}(S_1) + \text{Var}(S_2) + \text{Var}(S_3)$$

$$= 25 + 3(4)$$

$$= 37$$

$$\therefore V \sim N(5, 37)$$

$$P(E < S_1 + S_2 + S_3) = P(E - (S_1 + S_2 + S_3) < 0)$$

$$= P(V < 0)$$

$$\approx 0.206$$

- 18 a i Let X = the number of people cured then $X \sim B(100, \frac{3}{4})$.

ii $\mu_X = np$ $\sigma_X = \sqrt{np(1-p)}$
 $= 100 \times \frac{3}{4}$ $= \sqrt{100 \times \frac{3}{4} \times \frac{1}{4}}$
 $= 75$ ≈ 4.3301

iii $P(X \leq 68) \approx 0.0693$

iv $\bar{X} \sim N(0.75, \frac{4.3301^2}{100})$ {CLT}

$$\therefore P(\bar{X} \leq 0.68) \approx 0.0530$$

- b From a iv, the probability of getting 68 cured patients in a sample of 100 is very low (about 5.3%). This suggests that either the sample was biased or the company's claim of a 75% cure rate is not justified.

EXERCISE E.3

- 1 a Claim is: $p = 0.04$ and $n = 1000$

As $np = 40$ and $n(1-p) = 960$ are both ≥ 10 we can

$$\text{assume that } \hat{p} \sim N\left(0.04, \frac{0.04 \times 0.96}{1000}\right)$$

$$P(\hat{p} \geq 0.07) \approx 6.46 \times 10^{-7}$$

- b With such a small probability we would reject the egg producers claim.

- 2 $p = \frac{2}{7}$, $n = 100$

As $np \approx 28.57$ and $n(1-p) \approx 71.43$ are both ≥ 10 we can

$$\text{assume that } \hat{p} \sim N\left(\frac{2}{7}, \frac{\frac{2}{7} \times \frac{5}{7}}{100}\right)$$

$$\therefore \hat{p} \sim N\left(\frac{2}{7}, \frac{1}{490}\right)$$

$$\text{and } P(\hat{p} < \frac{29}{100}) \approx 0.538$$

- 3 a i $p = 0.465$ and $n = 2500$

$$\therefore \mu_{\hat{p}} = p = 0.465$$

ii $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.465 \times 0.535}{2500}}$
 ≈ 0.00998

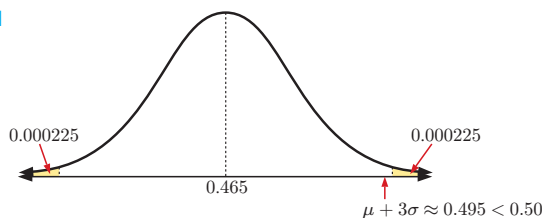
- b $\hat{p} \sim B(2500, 0.465)$ and is approximated by $N(0.465, 0.00998^2)$

- c i $P(\hat{p} < 0.46) \approx 0.308$

ii $P(0.45 < \hat{p} < 0.47) \approx 0.626$

iii $P(\hat{p} \text{ differs by more than } 3.5\% \text{ from } p)$
 $= P(\hat{p} < 0.43 \text{ or } \hat{p} > 0.50)$
 $= 1 - P(0.43 \leq \hat{p} \leq 0.50)$
 ≈ 0.000451

d



$\hat{p} \pm 3.5\%$ lies outside the range $\mu \pm 3\sigma$, so this probability is extremely small.

- 4 a $p = 0.85$ and for n sufficiently large

$$\hat{p} \sim N\left(0.85, \frac{0.85 \times 0.15}{n}\right)$$

- b For $\hat{p}$ to be approximated by the normal distribution we require that

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$

$$\therefore 0.85n \geq 10 \quad \text{and} \quad 0.15n \geq 10$$

$$\therefore n \geq 11.76 \quad \text{and} \quad n \geq 66.67$$

$$\therefore n \geq 67$$

c i $\hat{p} \sim N\left(0.85, \frac{0.85 \times 0.15}{200}\right)$

$$\therefore P(\hat{p} < 0.75) \approx 0.0000374$$

ii $P(0.75 < \hat{p} < 0.87) \approx 0.786$

- d i $n = 500$, $\hat{p} = \frac{350}{500} = 0.7$ and $p = 0.85$

$$np = 425 \text{ and } n(1-p) = 75 \text{ are both } \geq 10$$

$$\text{Now } \hat{p} \sim N\left(0.85, \frac{0.85 \times 0.15}{500}\right)$$

$$\therefore P(\hat{p} \leq \frac{350}{500}) \approx 0 \quad \{\text{using technology}\}$$

- ii Under the given conditions there is virtually no chance of this happening.

This means that either:

- (1) it was a freak occurrence which is possible but extremely unlikely
- (2) the population proportion was no longer 85% (probably < 85%)
- (3) the sample was not taken from the area mentioned.

$$5 \quad n = 400, \quad \hat{p} \sim N\left(\frac{2}{5}, \frac{\frac{2}{5} \times \frac{3}{5}}{400}\right)$$

$$\therefore \hat{p} \sim N(0.4, 0.0006)$$

$$a \quad P(\hat{p} > \frac{150}{400}) \approx 0.846 \quad b \quad P(\hat{p} < \frac{175}{400}) \approx 0.937$$

- 6 a $n = 250$ and their claim is $p = 0.9$
 $np = 250 \times 0.9 = 225$ and $n(1-p) = 25$ and these are both ≥ 10

$$\therefore \hat{p} \sim N\left(0.9, \frac{0.9 \times 0.1}{250}\right)$$

$$\therefore \hat{p} \sim N(0.9, 0.00036)$$

Assumptions made are:

- (1) the approximation to normal is satisfactory
- (2) When selected at random, the life of any tyre is independent of the life of any other tyre.

$$b \quad P(\hat{p} \leq \frac{200}{250})$$

$$= P(\hat{p} \leq 0.8)$$

$$\approx 6.82 \times 10^{-8} \quad \text{which is almost zero}$$

$\therefore$ the chance that $\hat{p} \leq 0.8$ is virtually impossible.

- c As the chance is practically zero there is little chance that the manufacturers claim is correct.

EXERCISE F

$$1 \quad a \quad T_1 = \frac{4}{12}X_1 + \frac{3}{12}X_2 + \frac{5}{12}X_3$$

$$\therefore E(T_1) = \frac{4}{12}E(X_1) + \frac{3}{12}E(X_2) + \frac{5}{12}E(X_3)$$

$$= \frac{4}{12}\mu + \frac{3}{12}\mu + \frac{5}{12}\mu$$

$$= \mu$$

$\therefore T_1$ is an unbiased estimator of μ .

$$b \quad T_2 = \frac{2}{6}X_1 + \frac{1}{6}X_2 + \frac{3}{6}X_3$$

$$\therefore E(T_2) = \frac{2}{6}E(X_1) + \frac{1}{6}E(X_2) + \frac{3}{6}E(X_3)$$

$$= \frac{2}{6}\mu + \frac{1}{6}\mu + \frac{3}{6}\mu$$

$$= \mu$$

$\therefore T_2$ is an unbiased estimator of μ .

$$c \quad \text{Var}(T_1) = \left(\frac{4}{12}\right)^2 \text{Var}(X_1) + \left(\frac{3}{12}\right)^2 \text{Var}(X_2) + \left(\frac{5}{12}\right)^2 \text{Var}(X_3)$$

$$= \frac{16}{144}\sigma^2 + \frac{9}{144}\sigma^2 + \frac{25}{144}\sigma^2$$

$$= \frac{50}{144}\sigma^2$$

$$\text{Var}(T_2) = \left(\frac{2}{6}\right)^2 \sigma^2 + \left(\frac{1}{6}\right)^2 \sigma^2 + \left(\frac{3}{6}\right)^2 \sigma^2$$

$$= \frac{14}{36}\sigma^2$$

$$\left(= \frac{56}{144}\sigma^2\right)$$

$$\therefore \text{Var}(T_1) < \text{Var}(T_2)$$

$\Rightarrow T_1$ is a more efficient estimator of μ than T_2 .

- 2 a $\bar{x}_{10}$ has distribution

$$\bar{X}_{10} = \frac{X_1 + X_2 + X_3 + \dots + X_{10}}{10} \quad \text{with}$$

$$E(\bar{X}_{10}) = \mu \quad \text{and} \quad \text{Var}(\bar{X}_{10}) = \frac{\sigma^2}{10}$$

Likewise $\bar{x}_{25}$ has distribution

$$\bar{X}_{25} = \frac{X_1 + X_2 + \dots + X_{25}}{25} \quad \text{with} \quad E(\bar{X}_{25}) = \mu \quad \text{and}$$

$$\text{Var}(\bar{X}_{25}) = \frac{\sigma^2}{25} \quad \text{and so} \quad \text{Var}(\bar{X}_{25}) < \text{Var}(\bar{X}_{10})$$

$\Rightarrow \bar{X}_{25}$ is a more efficient estimator of μ and so is preferred.

- b The larger the sample size n , the smaller the value of

$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Thus, as n increases, the more efficient the estimator $\bar{X}$ is of μ .

- 3 a $E(T_1) = E(T_2) = \theta$

$$E(T) = aE(T_1) + bE(T_2)$$

$$= a\theta + b\theta$$

$$= (a+b)\theta$$

$$\text{and } E(T) = \theta \Leftrightarrow a+b=1$$

$$b \quad T = \sum_{i=1}^n a_i T_i$$

$$\therefore E(T) = \sum_{i=1}^n a_i E(T_i)$$

$$= \sum_{i=1}^n a_i \theta$$

$$= \theta \sum_{i=1}^n a_i$$

$$\text{and } E(T) = \theta \Leftrightarrow \sum_{i=1}^n a_i = 1$$

- 4 If $T = \lambda X_1 + (1-\lambda)X_2$, $0 \leq \lambda \leq 1$

$$a \quad E(T) = \lambda E(X_1) + (1-\lambda)E(X_2)$$

$$= \lambda\mu + (1-\lambda)\mu$$

$$= \mu$$

$\therefore T$ is an unbiased estimator for μ .

$$b \quad \text{Var}(T) = \lambda^2 \text{Var}(X_1) + (1-\lambda)^2 \text{Var}(X_2)$$

$$= \lambda^2 \sigma^2 + (1-\lambda)^2 \sigma^2$$

$$= (\lambda^2 + 1 - 2\lambda + \lambda^2) \sigma^2$$

$$= (2\lambda^2 - 2\lambda + 1) \sigma^2$$

To get the most efficient estimator of T of this form we need to minimise $\text{Var}(T)$ and hence $2\lambda^2 - 2\lambda + 1$ (as σ^2 is constant).

$$\text{Consider } f(\lambda) = 2\lambda^2 - 2\lambda + 1$$

$$f'(\lambda) = 4\lambda - 2$$

$$\therefore f'(\lambda) = 0 \Leftrightarrow \lambda = \frac{1}{2}$$

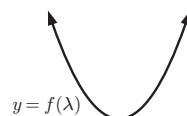
$$\therefore f(\lambda) \text{ is least when } \lambda = \frac{1}{2}$$

$$\therefore \text{Var}(T) \text{ is least when } \lambda = \frac{1}{2}$$

$\therefore$ the most efficient estimator of T has $\lambda = \frac{1}{2}$ and is

$$T = \frac{X_1 + X_2}{2}$$

= the sample mean



$$\begin{aligned}
 5 \quad a \quad E(T) &= E\left(\frac{3}{9}S_X^2 + \frac{6}{9}S_Y^2\right) \\
 &= \frac{3}{9}E(S_X^2) + \frac{6}{9}E(S_Y^2) \\
 &= \frac{3}{9}E(S_{4-1}^2) + \frac{6}{9}E(S_{7-1}^2) \\
 &= \frac{3}{9}\sigma^2 + \frac{6}{9}\sigma^2 \\
 &= \sigma^2
 \end{aligned}$$

$$\begin{aligned}
 b \quad \text{Let } t &= \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} \\
 \therefore E(T) &= \frac{1}{n+m-2}[(n-1)E(S_X^2) + (m-1)E(S_Y^2)] \\
 &= \frac{1}{n+m-2}[(n-1)\sigma^2 + (m-1)\sigma^2] \\
 &\quad \{\text{as } S_X^2 \text{ and } S_Y^2 \text{ are unbiased estimates of } \sigma^2\} \\
 &= \frac{1}{\cancel{n+m-2}} \sigma^2 (\cancel{n+m-2}) \\
 &= \sigma^2
 \end{aligned}$$

$$\begin{aligned}
 6 \quad a \quad X &\sim N(\mu, \sigma^2) \\
 \bar{X} &= \frac{X_1 + X_2 + X_3 + \dots + X_n}{n} \\
 \therefore E(\bar{X}) &= \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \dots + \frac{1}{n}E(X_n) \\
 &= \frac{1}{n}(\mu + \mu + \dots + \mu) \quad \{n \text{ } \mu\text{'s}\} \\
 &= \frac{1}{n}n\mu \\
 &= \mu \\
 \text{Var}(\bar{X}) &= \frac{1}{n^2}\text{Var}(X_1) + \frac{1}{n^2}\text{Var}(X_2) + \dots + \frac{1}{n^2}\text{Var}(X_n) \\
 &= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \dots + \sigma^2) \\
 &= \frac{n\sigma^2}{n^2} \\
 &= \frac{\sigma^2}{n}
 \end{aligned}$$

$$\text{But } \text{Var}(\bar{X}) = E(\bar{X}^2) - [E(\bar{X})]^2$$

$$\therefore \frac{\sigma^2}{n} = E(\bar{X}^2) - \mu^2$$

$$\therefore E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n} \quad \text{where } \frac{\sigma^2}{n} > 0$$

$$\therefore E(\bar{X}^2) > \mu^2$$

$$b \quad \text{As } E(\bar{X}^2) \neq \mu^2, \bar{X}^2 \text{ is a biased estimator of } \mu^2.$$

$$\begin{aligned}
 7 \quad E(X) &= \mu_X, \quad \text{Var}(X) = \sigma_X^2 \\
 E(Y) &= \mu_Y, \quad \text{Var}(Y) = \sigma_Y^2
 \end{aligned}$$

$$\begin{aligned}
 a \quad i \quad U &= X + Y \\
 E(U) &= E(X) + E(Y) = \mu_X + \mu_Y \\
 \text{and } \text{Var}(U) &= \text{Var}(X) + \text{Var}(Y) \\
 &= \sigma_X^2 + \sigma_Y^2
 \end{aligned}$$

$$\begin{aligned}
 ii \quad E(\bar{X} + \bar{Y}) &= E(\bar{X}) + E(\bar{Y}) \\
 &= \mu_X + \mu_Y \\
 &= E(U)
 \end{aligned}$$

$$\therefore \bar{x} + \bar{y} \text{ is an unbiased estimate of } E(U).$$

$$\begin{aligned}
 iii \quad E(S_X^2 + S_Y^2) &= E(S_X^2) + E(S_Y^2) \\
 &= \text{Var}(X) + \text{Var}(Y) \\
 &= \text{Var}(X + Y) \\
 &= \text{Var}(U) \\
 \therefore s_X^2 + s_Y^2 &\text{ is an unbiased estimate of } \text{Var}(U).
 \end{aligned}$$

$$b \quad U = aX + bY, \quad a, b \in \mathbb{R}^+$$

$$\begin{aligned}
 i \quad E(X) &= E(aX + bY) \\
 &= aE(X) + bE(Y) \\
 &= a\mu_X + b\mu_Y \\
 \text{Var}(U) &= \text{Var}(aX + bY) \\
 &= a^2\text{Var}(X) + b^2\text{Var}(Y) \\
 &= a^2\sigma_X^2 + b^2\sigma_Y^2
 \end{aligned}$$

$$\begin{aligned}
 ii \quad E(a\bar{X} + b\bar{Y}) &= aE(\bar{X}) + bE(\bar{Y}) \\
 &= a\mu_X + b\mu_Y \\
 &= E(U)
 \end{aligned}$$

$$\therefore a\bar{x} + b\bar{y} \text{ is an unbiased estimator for } E(U)$$

$$\therefore a\bar{x} + b\bar{y} \text{ is an unbiased estimate for } E(U)$$

$$\begin{aligned}
 iii \quad E(aS_X^2 + bS_Y^2) &= aE(S_X^2) + bE(S_Y^2) \\
 &= a\text{Var}(X) + b\text{Var}(Y) \\
 &\neq \text{Var}(U) \quad \{\text{from } b \text{ i}\} \\
 \therefore as_X^2 + bs_Y^2 &\text{ is generally not an unbiased estimate} \\
 &\text{of } \text{Var}(U).
 \end{aligned}$$

$$8 \quad \text{For the sample } \bar{x} = 4.39$$

$$s_X \approx 2.79358 = s_{n-1}$$

$$\sigma_X \approx 2.55018 = s_n$$

$$a \quad \text{An unbiased estimate of } \mu \text{ is } \bar{x} = 4.39$$

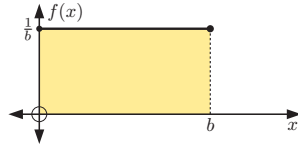
$$b \quad \text{An unbiased estimate of } \sigma^2 \text{ is } \approx 2.79358^2 \approx 7.8041$$

$$c \quad s_n^2 \approx (2.55018)^2 \approx 6.5034$$

$$d \quad s_{n-1}^2 = \frac{n}{n-1}s_n^2 = \frac{6}{5}s_n^2$$

$$\text{Check: } \frac{s_{n-1}^2}{s_n^2} \approx 1.2 \quad \checkmark$$

$$9 \quad X \sim U(0, b)$$



$$\begin{aligned}
 a \quad E(X) &= \int_0^b x f(x) dx = \int_0^b \frac{1}{b} x dx \\
 &= \left[\frac{1}{b} \frac{x^2}{2} \right]_0^b \\
 &= \frac{b}{2}
 \end{aligned}$$

$$b \quad \bar{X} \text{ is a sample mean estimator of } \frac{b}{2}$$

$$\text{Now } E(2\bar{X}) = 2E(\bar{X}) = 2\left(\frac{b}{2}\right) = b$$

$$\therefore 2\bar{x} \text{ is an unbiased estimator of } b \text{ for all } n \in \mathbb{Z}^+.$$

$$\begin{aligned}
10 \quad S_\mu^2 &= \frac{1}{n} \left[\sum_{i=1}^n (X_i - \mu)^2 \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^n (X_i^2 - 2\mu X_i + \mu^2) \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^n X_i^2 - 2\mu \sum_{i=1}^n X_i + n\mu^2 \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^n X_i^2 - 2\mu(n\mu) + n\mu^2 \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^n X_i^2 - n\mu^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^n X_i^2 - \mu^2 \\
\therefore E(S_\mu^2) &= \frac{1}{n} E \left(\sum_{i=1}^n X_i^2 \right) - \mu^2 \\
&= \frac{1}{n} \left[\sum_{i=1}^n E(X_i^2) \right] - \mu^2 \\
&= \frac{1}{n} \left[\sum_{i=1}^n (\text{Var}(X_i) + [E(X_i)]^2) \right] - \mu^2 \\
&\quad \{ \text{Var}(X) = E(X^2) - [E(X)]^2 \} \\
&= \frac{1}{n} (n\sigma^2 + n\mu^2) - \mu^2 \\
&= \sigma^2 + \mu^2 - \mu^2 \\
&= \sigma^2 \\
\therefore S_\mu^2 &\text{ is an unbiased estimator of } \sigma^2.
\end{aligned}$$

$$\begin{aligned}
11 \quad \text{a} \quad i \quad t &= \frac{4(3) + 9(5) + 20(2)}{30} = \frac{97}{30} \\
\text{ii} \quad \text{For sample A, } s_{n-1}^2 &= \frac{4}{3} s_n^2 \\
&\quad \therefore s_n^2 = \frac{3}{4} s_{n-1}^2 \\
\therefore E(S_A^2) &= \frac{3}{4} E(S_{n-1}^2) = \frac{3}{4} \sigma^2 \\
\{ \text{so } s_{n-1}^2 &\text{ is an unbiased estimate of } \sigma^2 \}
\end{aligned}$$

Likewise for sample B,

$$E(S_B^2) = \frac{8}{9} E(S_{n-1}^2) = \frac{8}{9} \sigma^2$$

For sample C,

$$\begin{aligned}
E(S_C^2) &= \frac{19}{20} E(S_{n-1}^2) = \frac{19}{20} \sigma^2 \\
\therefore E(T) &= \frac{4}{30} E(S_A^2) + \frac{9}{30} E(S_B^2) + \frac{20}{30} E(S_C^2) \\
&= \frac{4}{30} \left(\frac{3}{4} \sigma^2 \right) + \frac{9}{30} \left(\frac{8}{9} \sigma^2 \right) + \frac{20}{30} \left(\frac{19}{20} \sigma^2 \right) \\
&= \left(\frac{3 + 8 + 19}{30} \right) \sigma^2 \\
&= \sigma^2
\end{aligned}$$

$\therefore t$ is an unbiased estimate of σ^2 .

$$\begin{aligned}
b \quad t &= \frac{n_1 s_1^2 + n_2 s_2^2 + n_3 s_3^2 + \dots + n_r s_r^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + \dots + (n_r - 1)} \\
\text{or} \quad t &= \frac{\sum_{i=1}^r n_i s_i^2}{\left(\sum_{i=1}^r n_i \right) - r}
\end{aligned}$$

$$\begin{aligned}
12 \quad E(\bar{X}\bar{Y}) &= E \left[\frac{1}{n} \sum_{i=1}^n X_i \times \frac{1}{m} \sum_{j=1}^m Y_j \right] \\
&= \frac{1}{mn} E \left(\sum_{i=1}^n X_i \times \sum_{j=1}^m Y_j \right) \\
&= \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m E(X_i Y_j) \\
&= \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m E(X_i) E(Y_j) \quad \{ \text{as } X_i \text{ and } Y_j \text{ are independent} \} \\
&= \frac{1}{mn} \left(\sum_{i=1}^n E(X_i) \right) \left(\sum_{j=1}^m E(Y_j) \right) \\
&= \frac{1}{mn} (n\mu_X)(m\mu_Y) \\
&= \mu_X \mu_Y \\
\therefore \bar{x}\bar{y} &\text{ is an unbiased estimate of } \mu_X \mu_Y.
\end{aligned}$$

Note: In order to see this argument clearly work it through with $x = 2$, $m = 3$ say.

$$\begin{aligned}
13 \quad \text{For } n \text{ sufficiently large } \hat{p} &\sim N \left(p, \frac{p(1-p)}{n} \right) \\
\text{a} \quad E(\hat{p}) &= p \quad \text{and} \quad E(\hat{q}) = E(1 - \hat{p}) \\
&= 1 - E(\hat{p}) \\
&= 1 - p \\
\text{b} \quad E(\hat{p}\hat{q}) &= E(\hat{p}(1 - \hat{p})) \\
&= E(\hat{p} - \hat{p}^2) \\
&= E(\hat{p}) - E(\hat{p}^2) \\
\text{But } \text{Var}(\hat{p}) &= E(\hat{p}^2) - [E(\hat{p})]^2 \\
\therefore E \left(\frac{\hat{p}\hat{q}}{n} \right) &= \frac{1}{n} [E(\hat{p}) - \{ \text{Var}(\hat{p}) + [E(\hat{p})]^2 \}] \\
&= \frac{1}{n} \left[p - \frac{p(1-p)}{n} - p^2 \right] \\
&= \frac{np - p + p^2 - np^2}{n^2} \\
&= \frac{(n-1)p - (n-1)p^2}{n^2} \\
&= \frac{n-1}{n} \left(\frac{p(1-p)}{n} \right) \\
&= \left(\frac{n-1}{n} \right) \sigma_p^2 \\
\text{c} \quad \text{Since } E \left(\frac{\hat{p}\hat{q}}{n} \right) &\neq \sigma_p^2, \quad \frac{\hat{p}\hat{q}}{n} \text{ is a biased estimate of } \sigma_p^2. \\
\text{d} \quad \text{As } \left(\frac{n}{n-1} \right) \frac{\hat{p}\hat{q}}{n} &= \frac{\hat{p}\hat{q}}{n-1}, \\
E \left(\frac{\hat{p}\hat{q}}{n-1} \right) &= \frac{n}{n-1} E \left(\frac{\hat{p}\hat{q}}{n} \right) \\
&= \frac{n}{n-1} \left(\frac{n-1}{n} \right) \sigma_p^2 \\
&= \sigma_p^2 \\
\therefore \frac{\hat{p}\hat{q}}{n-1} &\text{ is an unbiased estimate of } \sigma_p^2.
\end{aligned}$$

EXERCISE G.1

- 1 μ is unknown, $\sigma = 10$, $n = 35$, $\bar{x} = 28.9$

a Using a Z -distribution, the 95% confidence interval is

$$\bar{x} - 1.960 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.960 \frac{\sigma}{\sqrt{n}}$$

Using technology, $25.6 < \mu < 32.2$

b Using a Z -distribution, the 99% confidence interval is

$$\bar{x} - 2.576 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.576 \frac{\sigma}{\sqrt{n}}$$

Using technology, $24.5 < \mu < 33.3$

c The confidence interval width becomes larger.

- 2 When increasing the level of accuracy we increase the interval width (as shown in question 1). We can estimate μ in a narrower interval but with less certainty.

- 3 Sample size n , $\sigma = 11$, $\bar{x} = 81.6$

a i $n = 36$, Z -distribution

$$\bar{x} - 1.960 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.960 \frac{\sigma}{\sqrt{n}}$$

Using technology, $78.0 < \mu < 85.2$

ii $n = 100$

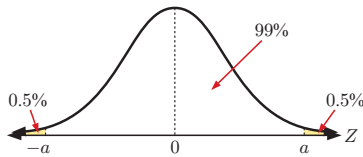
Using technology, $79.4 < \mu < 83.8$

b As n increases, the width of the confidence interval decreases.

- 4 Using the Z -distribution, the 95% confidence interval for μ is

$$\bar{x} - 1.960 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.960 \frac{\sigma}{\sqrt{n}} \quad \text{where } a = 1.960.$$

a If $P = 99$, we need to solve $P(Z < a) = 0.995$



Using technology gives $a \approx 2.576$

b If $P = 80$, we need to solve $P(Z < a) = 0.9$

Using technology gives $a \approx 1.282$

c If $P = 85$, we need to solve $P(Z < a) = 0.925$

$\therefore a \approx 1.440$

d If $P = 96$, we need to solve $P(Z < a) = 0.98$

$\therefore a \approx 2.054$

- 5 $n = 50$, standard deviation $= \sigma$, $\bar{x} = 38.7$, Z -distribution 95% confidence interval for μ is $\bar{x} - 1.960 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.960 \frac{\sigma}{\sqrt{n}}$

a For $\sigma = 6$, using technology, $37.0 < \mu < 40.4$

b For $\sigma = 15$, $34.5 < \mu < 42.9$

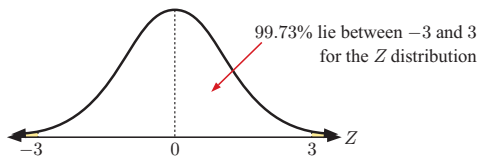
c As σ increases, the width increases.

- 6 $n = 167$, $\bar{x} = 8.7$, $2.6 \leq X \leq 15.1$

a range $= 15.1 - 2.6$

$= 12.5$

and $\sigma \approx \text{range} \div 6 \approx 2.083$



So, nearly all scores lie in the interval $[-3, 3]$ which has length 6 standard deviations.

- b 98% confidence interval for μ is

$$\bar{x} - 2.326 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.326 \frac{\sigma}{\sqrt{n}}$$

Using technology this is $8.33 < \mu < 9.08$

- 7 $n = 75$, $\bar{x} = 513.8$, $s_n = 14.9$

$$s_{n-1} = \sqrt{\frac{n}{n-1}} s_n = \sqrt{\frac{75}{74}} \times 14.9$$

$\therefore s_{n-1} \approx 15.00$ {unbiased estimate of σ }

As we had to estimate σ using s_n from the sample, the t -distribution applies.

The 99% confidence interval is $509.2 < \mu < 518.4$

- 8 $n = 42$, $\bar{x} = 38.2$, $s_n = 4.7$

$$s_{n-1} = \sqrt{\frac{n}{n-1}} s_n = \sqrt{\frac{42}{41}} \times 4.7$$

$\therefore s_{n-1} \approx 4.757$ {unbiased estimate of σ }

As we had to estimate σ from s_n , the t -distribution applies.

The 90% confidence interval is $37.0 < \mu < 39.4$

- 9 $n = 60$, $\bar{x} = 84.6$, $s_n = 16.8$

$$s_{n-1} = \sqrt{\frac{n}{n-1}} s_n = \sqrt{\frac{60}{59}} \times 16.8$$

$\therefore s_{n-1} \approx 16.94$ {unbiased estimate of σ }

As we had to estimate σ from s_n , the t -distribution applies.

a i 95% confidence interval is $80.2 < \mu < 89.0$

ii 99% confidence interval is $78.8 < \mu < 90.4$

b As $n = 50$, n is sufficiently large to use the normal confidence interval

$$\therefore 84.6 - 1.960 \frac{\sigma}{\sqrt{n}} < \mu < 84.6 + 1.960 \frac{\sigma}{\sqrt{n}}$$

$$\therefore -1.960 \frac{\sigma}{\sqrt{n}} < \mu - 84.6 < 1.960 \frac{\sigma}{\sqrt{n}}$$

$$\therefore |\mu - 84.6| < 1.960 \frac{\sigma}{\sqrt{n}}$$

$$\text{Thus } 1.960 \times \frac{16.94}{\sqrt{n}} < 5$$

$$\therefore \sqrt{n} > \frac{1.96 \times 16.94}{5}$$

$$\therefore n > 44.1$$

$\therefore$ a sample of 45 or more is needed.

- 10 $\sum x = 112.5$ and $\sum x^2 = 1325.31$

$$\begin{aligned} \text{a } \bar{x} &= \frac{\sum x}{n} \\ &= \frac{112.5}{10} \\ &= 11.25 \end{aligned}$$

$$\begin{aligned} \text{b } s_n^2 &= \frac{\sum x^2}{n} - \bar{x}^2 \\ &= \frac{1325.31}{10} - 11.25^2 \\ &= 5.9685 \end{aligned}$$

$$s_{n-1} = \sqrt{\frac{n}{n-1}} s_n$$

gives $s_{n-1} \approx 2.575$

and s_{n-1} is an unbiased estimate of σ .

c As s_{n-1} is used as an unbiased estimate of σ the t -distribution applies.

From technology, $9.96 < \mu < 12.7$

- 11 Using a Z -distribution with $\sigma = 17.8$, the 98% confidence interval for μ is $\bar{x} - 2.326 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.326 \frac{\sigma}{\sqrt{n}}$

$$\therefore |\mu - \bar{x}| < 2.326 \frac{\sigma}{\sqrt{n}}$$

$$\text{Hence } \frac{2.326 \times 17.8}{\sqrt{n}} < 3$$

$$\therefore \sqrt{n} > \frac{2.326 \times 17.8}{3}$$

$$\therefore n > 190.46 \dots$$

$\therefore$ should sample at least 191 packets.

- 12 a $n = 48$, $s_n^2 = 22.09$

$$\begin{aligned} s_{n-1} &= \sqrt{\frac{n}{n-1}} s_n \\ &= \sqrt{\frac{48}{47}} \times 22.09 \\ &= 4.7497 \dots \\ &\approx 4.75 \end{aligned}$$

$\therefore s_{n-1} = 4.75$ is an unbiased estimate of σ .

- b As n is large the 99% confidence interval can be obtained using the Z -distribution.

The 99% confidence interval for μ is

$$\bar{x} - 2.576 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.576 \frac{\sigma}{\sqrt{n}}$$

$$\therefore |\mu - \bar{x}| < 2.576 \frac{\sigma}{\sqrt{n}}$$

$$\therefore \text{we require } \frac{2.576 \times 4.7497}{\sqrt{n}} < 1.8$$

$$\therefore \sqrt{n} > \frac{2.576 \times 4.7497}{1.8}$$

$$\therefore n > 46.20$$

$\therefore n$ should be at least 47.

EXERCISE G.2

- 1 Let X_1 = the throwing distance of a 13 year old
 X_2 = the throwing distance of the same 12 year old
 and consider $U = X_2 - X_1$.

Age	A	B	C	D	E	F	G	H	I	J	K
u	3	1	7	5	3	2	-1	11	6	5	4

- a $n = 11$, $\bar{u} = 4.182$ and $s_{n-1} = 3.219$
 σ is unknown, so we use the T -statistic

$$T = \frac{\bar{U} - \mu}{\frac{s_{n-1}}{\sqrt{11}}} \text{ is } T \sim t(10)$$

- i 95% confidence interval for μ is $2.02 \leq \mu \leq 6.34$

- ii 90% confidence interval for μ is $2.42 \leq \mu \leq 5.94$

- b The sample of 11 is extremely small and the mean improvement $\bar{u} < 5 \text{ km h}^{-1}$.
 There is insufficient evidence to accept the sports commission claim as values < 5 lies within both the 95% and 90% confidence intervals for μ .

Pair	A	B	C	D	E	F	G	H
d	0.2	0.6	-0.2	0.8	0.2	-0.2	0.5	0.2

- a $n = 8$, $\bar{d} = 0.2625$, $s_{n-1}^2 \approx 0.128393$

- b σ is unknown so we use the T -statistic

- i 95% confidence interval for μ is:

$$-0.0371 \leq \mu \leq 0.5621$$

- ii 99% confidence interval for μ is: $-0.181 \leq \mu \leq 0.706$

- c Both confidence intervals in b contain the value 0 and also negative values. So, it is possible that $\mu < 0$ at both 95% and 99% levels of confidence. That is, there is insufficient evidence at these levels to support the manufacturers claim.

- d At a level of confidence α , the confidence interval is:

$$\bar{d} - t_{\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{d} + t_{\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}}$$

$$\therefore \bar{d} + t_{\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}} \approx 0.557$$

$$\therefore t_{\frac{\alpha}{2}} = \frac{0.557 - 0.2625}{\frac{\sqrt{0.128393}}{\sqrt{8}}}$$

$$\therefore t_{\frac{\alpha}{2}} \approx 2.3247$$

$$\therefore \frac{\alpha}{2} \approx 0.026513$$

$$\therefore \alpha \approx 0.0530 \quad (5.3\%)$$

$\therefore$ we have a 94.7% confidence level.

EXERCISE H.1

- a A Type I error involves rejecting a true null hypothesis.

b A Type II error involves accepting a false null hypothesis.

c The null hypothesis is a statement of *no difference*.

d The alternative hypothesis is a statement that there is a difference.
- a i a Type I error ii a Type II error

b i a Type II error ii a Type I error
- a The alternative hypothesis (H_1) would be that the person on trial is guilty.

b a Type I error c a Type II error
- a A Type I error would result if X and Y are determined to have different effectiveness, when in fact they have the same.

b A Type II error would result if X and Y are determined to have the same effectiveness, when in fact they have different effectiveness.
- a H_0 : new globe has mean life 80 hours
 H_1 : new globe has mean life > 80 hours

b H_0 : new globe has mean life 80 hours
 H_1 : new globe has mean life < 80 hours
- H_0 : new design has top speed of 26.3 knots
 H_1 : new design has top speed > 26.3 knots

EXERCISE H.2

- $$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\therefore \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \text{ and } \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\therefore \bar{x} \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}} \text{ and } \bar{x} \geq \mu - 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\therefore \mu - 1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}$$
- a For $\alpha = 0.05$,
 $z_{\alpha} \approx 1.645$ and $z_{\frac{\alpha}{2}} \approx 1.960$

- b For $\alpha = 0.01$,
 $z_{\alpha} \approx 2.326$ and $z_{\frac{\alpha}{2}} \approx 2.576$

3 a $\sigma^2 = 15.79$, $n = 36$, $\bar{x} = 23.75$

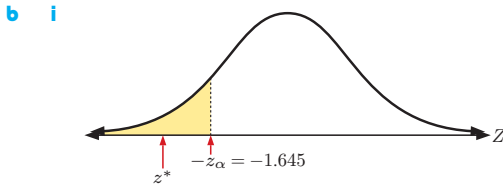
i
$$z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$\therefore z^* = \frac{23.75 - 25}{\frac{\sqrt{15.79}}{6}}$$

$$\therefore z^* \approx -1.887$$

ii The null distribution is $Z \sim N(0, 1)$

iii The p -value = $P(Z \leq -1.887)$
 ≈ 0.0296 {using technology}



As $z^* < -z_{\alpha}$ we reject H_0 in favour of H_1 .

ii As the p -value < 0.05 we reject H_0 in favour of H_1 .

4 a i $z > z_{\alpha}$, so $z > 1.645$

ii $z > z_{\alpha}$, so $z > 2.326$

b i $z < -z_{\alpha}$, so $z < -1.645$

ii $z < -z_{\alpha}$, so $z < -2.326$

c i $\frac{\alpha}{2} = 0.025$
 $z < -z_{\frac{\alpha}{2}}$ or $z > z_{\frac{\alpha}{2}}$
 $\therefore z < -1.960$ or $z > 1.960$

ii $\frac{\alpha}{2} = 0.005$
 $z < -z_{\frac{\alpha}{2}}$ or $z > z_{\frac{\alpha}{2}}$
 $\therefore z < -2.576$ or $z > 2.576$

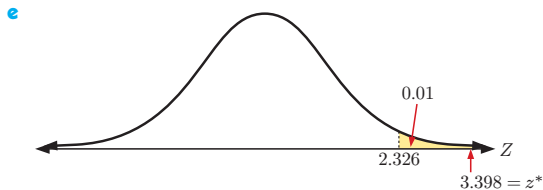
5 $\sigma = 12.9$, $n = 200$, $\bar{x} = 83.1$

a $H_0: \mu = 80$ and $H_1: \mu > 80$

b The null distribution is Z with $\sigma = 12.9$

c
$$z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{83.1 - 80}{\frac{12.9}{\sqrt{200}}} \approx 3.398$$

d Reject H_0 if either z^* lies in the critical region of Z or the p -value is < 0.01 .



f As z^* lies within the critical region we reject H_0 in favour of H_1 . or
 As the p -value = $P(Z \geq z^*) = 0.000339$ is < 0.01 we reject H_0 .

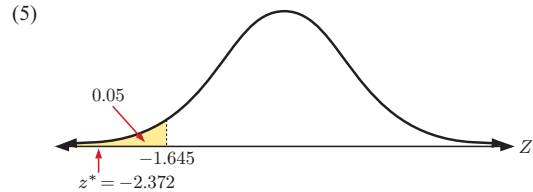
g We accept that $\mu > 80$ at the 1% level of significance. $P(\text{Type I error}) = 0.01$.

6 (1) $H_0: \mu = 100$ and $H_1: \mu < 100$

(2) As σ is known ($\sigma = 1.6$ g) the null distribution is Z .

(3) The test statistic is
$$z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{99.4 - 100}{\frac{1.6}{\sqrt{40}}} \approx -2.372$$

(4) We reject H_0 if z^* lies in the critical region.



(6) Since z^* lies in the critical region we reject H_0 in favour of H_1 .

(7) We have sufficient evidence to accept H_1 , that the mean weight is less than 100 g net. There is evidence that the machine which delivers the nuts needs to be adjusted to allow more nuts into each bag.

7 a (1) $H_0: \mu = 22.3$ (no difference)

$H_1: \mu \neq 22.3$ (a difference)

(2) Assuming $\sigma = 2.89$ is constant the null distribution is Z .

(3) The test statistic is
$$z^* = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{21.2 - 22.3}{\frac{2.89}{\sqrt{80}}} \approx -3.404$$

(4) We reject H_0 if the p -value < 0.05

(5) p -value = $P(Z < -3.404 \text{ or } Z > 3.404) \approx 0.000663$

(6) Since p -value < 0.05 we reject H_0 in favour of H_1 .

(7) There is sufficient evidence at a 5% level to suggest that the mean fleece diameter differs in 2012 from 2008.

b With $\bar{x} = 21.2$, $\sigma = 2.89$, $n = 80$

The 95% confidence interval is $20.57 < \mu < 21.83$ and $\mu_0 = 21.2$ lies within it.

This confirms, at a 95% level of confidence that there is a significant difference in the means between 2008 and 2012.

8 $\sigma^2 = 2.25$ is known, $n = 8$, $\bar{x} = 1001$

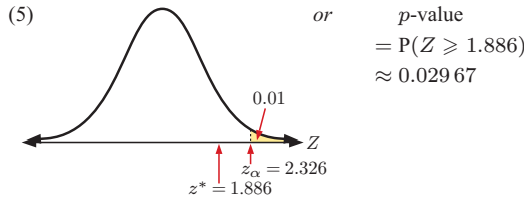
(1) $H_0: \mu = 1000$ g and $H_1: \mu > 1000$ g

(2) As σ^2 is known we use a Z -distribution.

(3) The test statistic is
$$z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{1001 - 1000}{\frac{\sqrt{2.25}}{\sqrt{8}}} = \frac{1 \times \sqrt{8}}{\sqrt{2.25}} \approx 1.886$$

(4) We reject H_0 if:

z^* lies in the critical region or the p -value < 0.01



- (6) As z^* is not in the critical region we do not reject H_0 . or As $p\text{-value}$ is not < 0.01 we do not reject H_0 .
- (7) We conclude that there is insufficient evidence to support the overfilling claim at a 1% level. (However we may be making a Type II error.)

EXERCISE H.3

- 1 a For $\nu = 15$ and $\alpha = 0.05$,
 $t_{\alpha} \approx 1.753$ and $t_{\frac{\alpha}{2}} \approx 2.131$

- b For $\nu = 15$ and $\alpha = 0.01$,
 $t_{\alpha} \approx 2.602$ and $t_{\frac{\alpha}{2}} \approx 2.947$

- 2 $H_0: \mu = 18.5$ and $H_1: \mu \neq 18.5$

- a i $n = 24$, $\bar{x} = 17.14$, $s_n = 4.365$

$$\begin{aligned} \text{Now } s_{n-1} &= \sqrt{\frac{n}{n-1}} s_n \\ &= \sqrt{\frac{24}{23}} \times 4.365 \\ &\approx 4.459 \end{aligned}$$

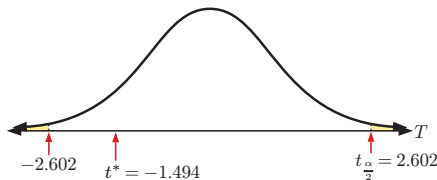
which is an unbiased estimate of σ and

$$t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \approx \frac{17.14 - 18.5}{\frac{4.459}{\sqrt{24}}}$$

$$\therefore t^* \approx -1.494$$

- ii The null distribution is $t(23)$.
- iii The $p\text{-value} = 2 \times P(T \geq |t^*|)$
 $= 2 \times P(T \geq 1.494)$
 ≈ 0.1487

- b i



As t^* does not lie in the critical region we do not reject H_0 as there is insufficient evidence to do so. We accept at a 5% level that $\mu = 18.5$.

- ii As $p\text{-value}$ is not < 0.05 we do not reject H_0 as there is insufficient evidence to do so. We accept at a 5% level that $\mu = 18.5$.

- 3 a $H_0: \mu = \$13.45$, $H_1: \mu < \$13.45$

- b As σ is unknown we use a t -distribution s_{n-1} is used as an unbiased estimate of σ and

$$\begin{aligned} s_{n-1} &= \sqrt{\frac{n}{n-1}} s_n \\ &= \sqrt{\frac{389}{388}} \times \$0.25 \\ &\approx \$0.2503 \end{aligned}$$

- c $t^* \approx -11.82$

- d $p\text{-value} = P(T < -11.82) = 0$ which is < 0.02

- e We reject H_0 that $\mu = \$13.45$ at a 2% level. That is, we accept the claim that the mean price has fallen.

Note: $P(\text{Type I error}) = 0.02$.

- 4 $\bar{x} = 499$ mL and $s_n = 1.2$ mL

As σ is unknown, we use s_{n-1} as an unbiased estimate of σ

$$\begin{aligned} \text{and } s_{n-1} &= \sqrt{\frac{n}{n-1}} s_n \\ &= \sqrt{\frac{10}{9}} \times 1.2 \\ &\approx 1.2649 \end{aligned}$$

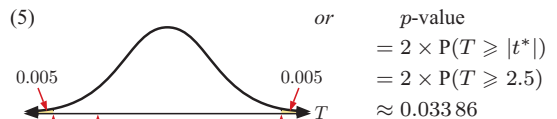
- (1) $H_0: \mu = 500$ mL and $H_1: \mu \neq 500$ mL

- (2) As σ is unknown we use the t -statistic, $t(9)$.

$$(3) \quad t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} = \frac{499 - 500}{\frac{1.2649}{\sqrt{10}}}$$

$$\therefore t^* \approx -2.500 \text{ is the test statistic}$$

- (4) We reject H_0 if:
 t^* lies in the critical region or the $p\text{-value} < 0.01$



- (6) As t^* is not in the critical region we do not reject H_0 . or As the $p\text{-value}$ is not < 0.01 we do not reject H_0 .

- (7) We conclude that there is insufficient evidence to suggest that the sample mean is significantly different from the expected value, at a 1% level. (However, we risk making a Type II error by accepting H_0 when it is false.)

- 5 a $n = 60$, $\bar{x} = 242.6$ mg, $s_n = 7.3$ mg

As σ is unknown, we use s_{n-1} as an unbiased estimate of

$$\begin{aligned} \sigma \text{ and } s_{n-1} &= \sqrt{\frac{n}{n-1}} s_n \\ &= \sqrt{\frac{60}{59}} \times 7.3 \\ &\approx 7.3616 \end{aligned}$$

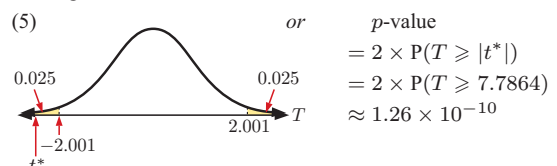
- (1) $H_0: \mu = 250$ mg and $H_1: \mu \neq 250$ mg

- (2) As σ is unknown we use the t -statistic $t(59)$.

$$(3) \quad t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \approx \frac{242.6 - 250}{\frac{7.3616}{\sqrt{60}}}$$

$$\therefore t^* \approx -7.7864$$

- (4) We reject H_0 if:
 t^* lies in the critical region or the $p\text{-value} < 0.05$



- (6) As t^* lies in the critical region we reject H_0 .
 or As the p -value < 0.05 we reject H_0 .
- (7) There is sufficient evidence to reject H_0 in favour of H_1 .
 This suggests that at a 5% level we should accept that $\mu \neq 250$ and as $\bar{x}$ was < 250 we surmise that the true μ is smaller than 250 mg.

b The 95% confidence interval for μ is $240.7 < \mu < 244.5$ which confirms the above as we are 95% confident that μ is well below 250 mg. Hence, we would reject H_0 in **a** and argue again that $\mu < 250$ mg.

6 $n = 50$, $\bar{x} = 26.1$, $s_n = 6.38$

As σ is unknown, we use s_{n-1} as an unbiased estimate of it

$$\text{and } s_{n-1} = \sqrt{\frac{n}{n-1}} s_n$$

$$= \sqrt{\frac{50}{49}} \times 6.38$$

$$\approx 6.4448$$

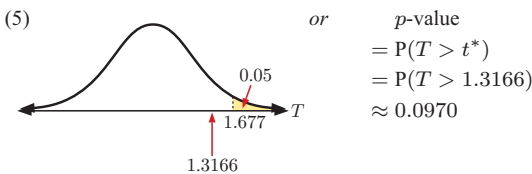
(1) $H_0: \mu = 24.9$ and $H_1: \mu > 24.9$

(2) As σ is unknown, we use the t -statistic, $t(49)$.

$$(3) \quad t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \approx \frac{26.1 - 24.9}{\frac{6.4448}{\sqrt{50}}}$$

$$\therefore t^* \approx 1.3166$$

(4) We reject H_0 if:
 t^* lies in the critical region or the p -value < 0.05



(6) As t^* is not in the critical region we do not reject H_0 .
 or As the p -value is not < 0.05 we do not reject H_0 .

(7) We conclude that, at a 5% level, there is insufficient evidence to suggest that the non-free range chickens have a greater meat protein content.

7 $\bar{x} = \text{€}96\,318$, $s_n = \text{€}14\,268$, $n = 113$

a $s_{n-1} = \sqrt{\frac{n}{n-1}} s_n$

$$= \sqrt{\frac{113}{112}} \times \text{€}14\,268$$

$$\approx \text{€}14\,331.55$$

$\therefore \text{€}14\,331.55$ is an unbiased estimate of σ .

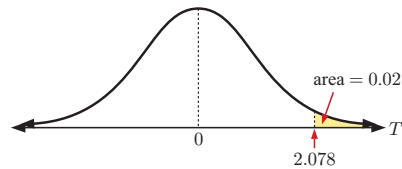
b $H_0: \mu = \text{€}95\,000$ and $H_1: \mu > \text{€}95\,000$

c As σ is unknown and we have to estimate it, the null distribution is a t -distribution with $\nu = n - 1 = 112$.

d $t^* \approx 0.9776$

e p -value $= P(t > 0.9776) \approx 0.1652$

f Critical value is $t_{0.02} \approx 2.078$ as we need to solve
 $P(t > k) = 0.02$
 or $P(t < k) = 0.98$
 using technology with **inverse t** .



g As $t^* \approx 0.9776$ is < 2.078 we have insufficient evidence to reject H_0 .

So, we reject the claim that $\mu > \text{€}95\,000$.

h If the assertion was incorrect, that is, accepting H_0 when H_1 is correct, we would be making a Type II error.

i The 99% confidence interval for the mean income is $\text{€}92\,785 < \mu < \text{€}99\,851$.

This interval confirms that there is not enough evidence to reject H_0 as $\mu = 95\,000$ lies within the interval.

Although $\alpha = 0.02$, we verify with a 99% confidence interval as we have a one-tailed test.

EXERCISE H.4

1 Let X_1 represent the test score before coaching,

X_2 represent the test score after coaching

and let $U = X_2 - X_1$.

U -values are 5, -1, 0, 7, 0, -1, 3, 3, 4, -1, 1, -6

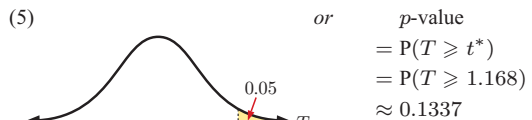
$$\bar{U} \approx 1.1667, \quad s_{n-1} \approx 3.4597, \quad n = 12$$

(1) $H_0: \mu = 0$ {no improvement}, $H_1: \mu > 0$

(2) As σ is not known we use a t -distribution with $t(11)$, and use s_{n-1} as an unbiased estimate of σ .

(3) Test statistic is $t^* \approx 1.168$

(4) We reject H_0 if:
 t^* does not lie in the critical region or the p -value is < 0.05



(6) As t^* does not lie in the rejection region we do not reject H_0 .
 or As the p -value is not < 0.05 we do not reject H_0 .

(7) We conclude that, at a 5% level, there is insufficient evidence to support that there has been improvement.

2 Let X_1 = speed at age 12,

X_2 = speed at age 13

and let $U = X_2 - X_1$, $n = 11$

The U -values are: 3, 1, 7, 5, 3, 2, -1, 11, 6, 5, 4

$$\bar{U} \approx 4.1818 \text{ and } s_{n-1} \approx 3.2193$$

(1) $H_0: \mu = 5$ and $H_1: \mu \neq 5$

(2) As σ is unknown we use the t -statistic, $t(10)$. We use s_{n-1} as an unbiased estimate of σ .

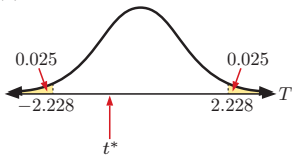
(3) The test statistic, $t^* = \frac{\bar{U} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}}$

$$\approx \frac{4.1818 - 5}{\frac{3.2193}{\sqrt{11}}}$$

$$\approx -0.84286$$

- (4) We reject H_0 if:
 t^* does not lie in the critical region or the p -value is < 0.05

(5) or p -value
 $= 2P(T \geq |t^*|)$
 $= 2 \times P(T \geq 0.84286)$
 ≈ 0.4190



- (6) As t^* does not lie in the critical region we do not reject H_0 . or As the p -value is not less than 0.05 we do not reject H_0 .
- (7) We conclude, at a 5% level, that there is insufficient evidence to reject the sports commission's claim.

- 3 Let X_1 = height using Type 1 compost,
 X_2 = height using Type 2 compost
 and let $U = X_2 - X_1$.

U -values are: 0.2, 0.6, -0.2, 0.8, 0.2, -0.2, 0.5, 0.2

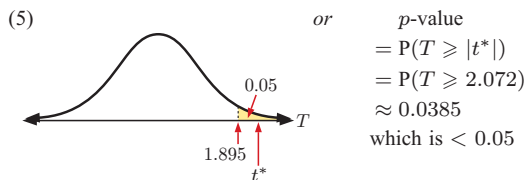
where $\bar{U} \approx 0.2625$ and $s_{n-1} \approx 0.35832$

- (1) $H_0: \mu = 0$ and $H_1: \mu > 0$
 (2) As σ^2 is unknown we use the t -statistic, $t(7)$ and s_{n-1} as an unbiased estimate of σ .

$$(3) \quad t^* = \frac{\bar{U} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \approx \frac{0.2625 - 0}{\frac{0.35832}{\sqrt{8}}}$$

$$\therefore t^* \approx 2.072$$

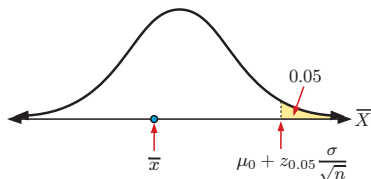
- (4) We reject H_0 if:
 t^* does not lie in the critical region or the p -value < 0.05



- (6) As t^* lies in the critical region we reject H_0 . or As the p -value is < 0.05 , we reject H_0 .
- (7) We conclude, at a 5% level, that there is sufficient evidence to support the claim that Type 2 compost improves growth over Type 1 compost.

EXERCISE H.5

1 a i



If $\alpha = 0.05$, we accept H_0 if

$$\bar{x} < \mu_0 + 1.6449 \frac{\sigma}{\sqrt{n}}$$

$$\therefore \bar{x} < 27 + 1.6449 \times \frac{\sqrt{6}}{\sqrt{9}}$$

$$\therefore \bar{x} < 28.343$$

- ii If $\alpha = 0.01$,
 we accept H_0 if $\bar{x} < 27 + 2.3263 \times \sqrt{\frac{2}{3}}$
 $\therefore \bar{x} < 28.899$

b i $P(\text{Type II error}), \alpha = 0.05$
 $= P(\text{accepting } H_0 \mid H_1 \text{ is true})$
 $= P(\bar{X} < 28.343 \mid \mu_0 = 29.2)$
 $= P(\bar{X} < 28.343 \mid \bar{X} \sim N(29.2, \frac{6}{9}))$
 ≈ 0.147

ii $P(\text{Type II error}), \alpha = 0.01$
 $= P(\text{accepting } H_0 \mid H_1 \text{ is true})$
 $= P(\bar{X} < 28.899 \mid \bar{X} \sim N(29.2, \frac{6}{9}))$
 ≈ 0.356

- 2 $n = 16, \mu = \text{unknown}, \sigma^2 = 64$
 $H_0: \mu = 150$ and $H_1: \mu > 150$
 Decision rule: accept H_0 if $\bar{x} \leq 155$ otherwise reject it.

a $\alpha = P(\text{Type I error})$
 $= P(\text{rejecting } H_0 \mid H_0 \text{ is true})$
 $= P(\bar{X} \geq 155 \mid \mu_0 = 150)$
 $= P(\bar{X} \geq 155 \mid \bar{X} \sim N(150, \frac{64}{16}))$
 $= P(\bar{X} \geq 155 \mid \bar{X} \sim N(150, 4))$
 ≈ 0.00621 (about 0.621%)

b i $P(\text{Type II error})$
 $= P(\text{accepting } H_0 \mid H_1 \text{ is true})$
 $= P(\bar{X} \leq 155 \mid \bar{X} \sim N(159, 4))$
 ≈ 0.0228

ii When $P(\text{Type I error}) = P(\text{Type II error})$,
 the critical value for $\bar{x} = \frac{150 + 159}{2}$
 $= 154.5$

Check:

$$P(\text{Type I error})$$

$$= P(\bar{X} \geq 154.5 \mid \bar{X} \sim N(150, 4))$$

$$\approx 0.01222$$

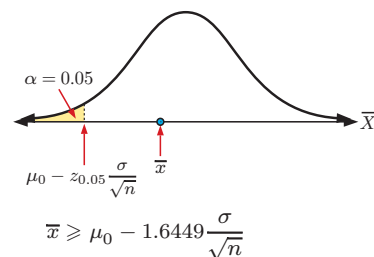
$$P(\text{Type II error})$$

$$= P(\bar{X} \leq 154.5 \mid \bar{X} \sim N(159, 4))$$

$$\approx 0.01222 \quad \checkmark$$

- 3 $n = 30, \mu = \text{unknown}, \sigma^2 = 7.5$
 $H_0: \mu = 37$ and $H_1: \mu < 37$

a



$$\bar{x} \geq \mu_0 - 1.6449 \frac{\sigma}{\sqrt{n}}$$

$$\therefore \bar{x} \geq 37 - 1.6449 \times \frac{\sqrt{7.5}}{\sqrt{30}}$$

$$\therefore \bar{x} \geq 37 - 1.6449 \times \sqrt{\frac{1}{4}}$$

$$\therefore \bar{x} \geq 36.178$$

$$\therefore \text{accept } H_0 \text{ if } \bar{x} \geq 36.178, \text{ otherwise reject it.}$$

- b** True mean is $\mu = 36$

$$\begin{aligned}\beta &= P(\text{Type II error}) \\ &= P(\text{accepting } H_0 \mid H_1 \text{ is true}) \\ &= P(\bar{X} \geq 36.178 \mid \bar{X} \sim N(36, \frac{1}{4})) \\ &\approx 0.361\end{aligned}$$

$$\begin{aligned}\therefore \text{power} &= 1 - \beta \approx 0.639 \\ &(\approx 63.9\%) \end{aligned}$$

- c** If $P(\text{Type II error}) = 0.1$

$$P(\text{accepting } H_0 \mid H_1 \text{ is true}) = 0.1$$

$$\therefore P(\bar{X} \geq 36.178 \mid \bar{X} \sim N(\mu, \frac{1}{4})) = 0.1$$

$$\therefore P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{36.178 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0.1$$

$$\therefore P\left(Z \geq \frac{36.178 - \mu}{0.5}\right) = 0.1$$

$$\therefore P\left(Z \leq \frac{36.178 - \mu}{0.5}\right) = 0.9$$

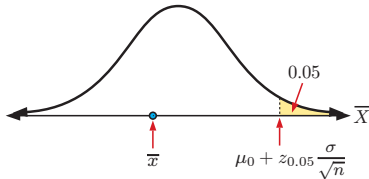
$$\therefore \frac{36.178 - \mu}{0.5} \approx 1.2816$$

$$\therefore 36.178 - \mu \approx 0.6408$$

$$\therefore \mu \approx 35.5$$

- 4** $\mu = 6.4$, $\sigma = 0.7$ kg, $n = 15$, $\alpha = 0.05$

- a** $H_0: \mu = 6$ and $H_1: \mu > 6$



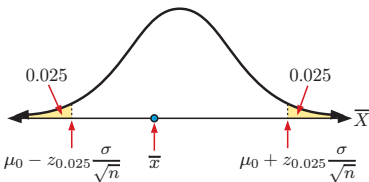
$$\text{Accept } H_0 \text{ if } \bar{x} \leq \mu_0 + 1.6449 \times \frac{0.7}{\sqrt{15}}$$

$$\therefore \bar{x} \leq 6.2973$$

otherwise reject it.

$$\begin{aligned}P(\text{Type II error}) &= P(\text{accepting } H_0 \mid H_1 \text{ is true}) \\ &= P(\bar{X} \leq 6.2973 \mid \bar{X} \sim (6.4, \frac{0.7^2}{15})) \\ &\approx 0.285\end{aligned}$$

- b** $H_0: \mu = 6$ and $H_1: \mu \neq 6$



$$\text{Accept } H_0 \text{ if } \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\therefore 6 - 1.960 \times \frac{0.7}{\sqrt{15}} \leq \bar{x} \leq 6 + 1.960 \times \frac{0.7}{\sqrt{15}}$$

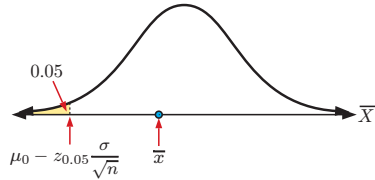
$$\therefore 5.6458 \leq \bar{x} \leq 6.3542,$$

otherwise reject it.

$$\begin{aligned}P(\text{Type II error}) &= P(\text{accepting } H_0 \mid H_1 \text{ is true}) \\ &= P(5.6458 \leq \bar{X} \leq 6.3542 \mid \bar{X} \sim N(6.4, \frac{0.7^2}{15})) \\ &\approx 0.400\end{aligned}$$

$$\therefore \text{power} \approx 1 - 0.400 \approx 0.600$$

- c** $H_0: \mu = 6$ and $H_1: \mu < 6$



$$\text{We accept } H_0 \text{ if } \bar{x} > 6 - 1.6449 \times \frac{0.7}{\sqrt{15}}$$

$$\therefore \text{accept } H_0 \text{ if } \bar{x} > 5.7027, \text{ otherwise reject it.}$$

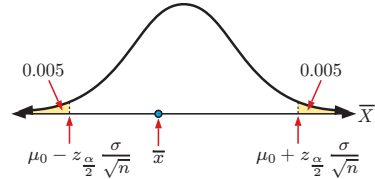
$$\begin{aligned}P(\text{Type II error}) &= P(\text{accepting } H_0 \mid H_1 \text{ is true}) \\ &= P(\bar{X} > 5.7027 \mid \bar{X} \sim N(6.4, \frac{0.7^2}{15})) \\ &\approx 0.99994 \\ \therefore \text{power} &\approx 0.00006 \text{ (virtually 0).}\end{aligned}$$

- 5** Let X = length of a beam (in m).

$$X \sim N(\mu, 0.15^2)$$

- a** $H_0: \mu = 3.5$ m and $H_1: \mu \neq 3.5$ m

b



We reject H_0 if

$$\mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\therefore 3.5 - 2.5758 \left(\frac{0.15}{\sqrt{20}}\right) \leq \bar{x} \leq 3.5 + 2.5758 \left(\frac{0.15}{\sqrt{20}}\right)$$

$$\therefore 3.4136 \leq \bar{x} \leq 3.5864$$

$$\begin{aligned}P(\text{Type II error}) &= P(\text{accepting } H_0 \mid H_1 \text{ is true}) \\ &= P(3.4136 \leq \bar{X} \leq 3.5864 \mid \bar{X} \sim N(3.4, \frac{0.15^2}{20})) \\ &\approx 0.343\end{aligned}$$

EXERCISE H.6

- 1 a** H_0 : The die is fair for rolling a '4', so $p = \frac{1}{4}$

$$H_1: \text{The die is unfair, so } p \neq \frac{1}{4}$$

- b i** A Type I error is rejecting H_0 when it is true. This means deciding it is biased when it is in fact fair.

$$\begin{aligned}\text{ii } P(\text{Type I error}) &= P(\text{Reject } H_0 \mid H_0 \text{ is true}) \\ &= P(X \leq 61 \text{ or } X \geq 89 \mid p = \frac{1}{4}) \\ &\quad \text{where } X \sim B(300, \frac{1}{4}) \\ &= 1 - P(62 \leq X \leq 88) \text{ with } X \sim B(300, \frac{1}{4}) \\ &= 1 - [P(X \leq 89) - P(X \leq 61)] \\ &\approx 1 - 0.97168 + 0.33785 \\ &\approx 0.0621\end{aligned}$$

- iii** The test is about the 6.2% level.

- c** If H_0 is true, $X \sim B(300, \frac{1}{4})$.

For the hypotheses in **a** we have a two-tailed test with

$$\alpha \approx 0.02 \text{ and } \frac{\alpha}{2} \approx 0.01.$$

Solving $P(X \leq k) = 0.01$ gives $k = 58$

$$P(X \leq k) = 0.99 \text{ gives } k = 92$$

The new decision rule is:

Roll the die 300 times. If the number of 4s obtained (X) is such that $58 \leq X \leq 92$, accept H_0 , otherwise reject it.

d If $p = 0.32$, then $X \sim b(300, 0.32)$

$\therefore$ P(Type II error)

$$= P(62 \leq X \leq 88 \mid X \sim B(300, 0.32))$$

$$= P(X \leq 88) - P(X \leq 61)$$

$$\approx 0.177$$

2 Y = volume of drink can (in cm^3)

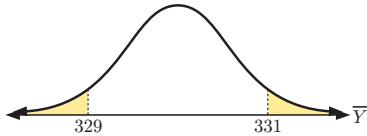
$$Y \sim N(\mu, 2^2)$$

$$\therefore \bar{Y} \sim N(\mu, \frac{2^2}{16})$$

$$\therefore \bar{Y} \sim N(\mu, (\frac{1}{2})^2)$$

$$H_0: \mu = 330 \text{ against } H_1: \mu \neq 330$$

Critical region is $\bar{y} < 329$ or $\bar{y} > 331$



a $\alpha = P(\text{Type I error})$

$$= P(\text{rejecting } H_0 \mid H_0 \text{ is true})$$

$$= P(\bar{Y} < 329 \text{ or } \bar{Y} > 331 \mid \mu = 330)$$

$$= 1 - P(329 \leq \bar{Y} \leq 331 \mid \mu = 330)$$

$$\approx 0.0455$$

So $\alpha \approx 4.55\%$

b P(Type II error)

$$= P(\text{accepting } H_0 \mid H_1 \text{ is true})$$

$$= P(329 \leq \bar{Y} \leq 331 \mid \bar{Y} \sim N(328, (\frac{1}{2})^2))$$

$$\approx 0.02275$$

3 a $H_0: p = 0.5$ (the coin is fair)

$H_1: p > 0.5$ (coin is biased towards heads)

b i $X \geq 10$ is the critical region.

ii The significance level (α) of a hypothesis test is the probability of making a Type I error. That is, the probability of rejecting the null hypothesis H_0 when it is indeed true.

$$\alpha = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$$

$$= P(X \geq 10 \mid X \sim B(12, \frac{1}{2}))$$

$$= 1 - P(X \leq 9 \mid X \sim B(12, 0.5))$$

$$\approx 0.0193$$

c If p is in fact 0.6

P(Type II error)

$$= P(\text{accepting } H_0 \mid p = 0.6)$$

$$= P(X \leq 9 \mid X \sim B(12, 0.6))$$

$$\approx 0.917$$

d Eri would accept H_0 when H_1 is true and so risks making a Type II error.

4 $H_0: \lambda = 2$ against $H_1: \lambda = 5$

We accept H_0 if $X \leq 3$

We reject H_0 if $X \geq 4$

a If $n = 1$, $X \sim \text{Po}(\lambda)$

i $P(\text{reject } H_0 \mid H_0 \text{ is true})$

$$= P(X \geq 4 \mid X \sim \text{Po}(2))$$

$$= 1 - P(X \leq 3 \mid X \sim \text{Po}(2))$$

$$\approx 0.143$$

ii $P(\text{accept } H_0 \mid H_1 \text{ is true})$

$$= P(X \leq 3 \mid X \sim \text{Po}(5))$$

$$\approx 0.265$$

b i We require that:

$$P(\text{accept } H_0 \mid H_1 \text{ is true}) \text{ to be } < 0.005$$

$$\therefore P(X \leq 3 \mid X \sim \text{Po}(5n)) < 0.005$$

$$\text{If } n = 1, \quad P(X \leq 3 \mid X \sim \text{Po}(5)) \approx 0.265 \quad \{\text{a ii}\}$$

$$\text{If } n = 2, \quad P(X \leq 3 \mid X \sim \text{Po}(10)) \approx 0.0103$$

$$\text{If } n = 3, \quad P(X \leq 3 \mid X \sim \text{Po}(15)) \approx 0.000211$$

which is < 0.005

$\therefore n = 3$ minutes is required.

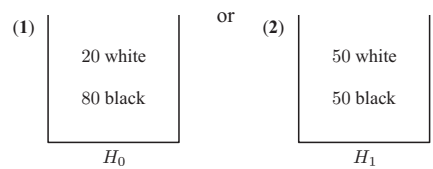
ii $P(\text{rejecting } H_0 \mid H_0 \text{ is true})$

$$= P(X \geq 4 \mid X \sim \text{Po}(3 \times 2))$$

$$= 1 - P(X \leq 3 \mid X \sim \text{Po}(6))$$

$$\approx 0.849$$

5 Either



We accept H_0 if all 4 are black

reject H_0 if some of the 4 are white.

a P(Type I error) = $P(\text{reject } H_0 \mid H_0 \text{ is true})$

$$= P(\text{not all 4 are black from box (1)})$$

$$= 1 - P(\text{all 4 are black from (1)})$$

$$= 1 - \frac{{}^8C_4 {}^0C_0}{{}^{100}C_4}$$

$$\approx 0.597$$

P(Type II error) = $P(\text{accepting } H_0 \mid H_1 \text{ is true})$

$$= P(\text{all 4 are black from box (2)})$$

$$= \frac{{}^5C_4 {}^50C_0}{{}^{100}C_4}$$

$$\approx 0.0587$$

b Power for above test = $1 - \beta$

$$\approx 1 - 0.0587$$

$$\approx 0.9413 \text{ or } 94.1\%$$

For the new test:

P(Type II error)

$$= P(\text{accepting } H_0 \mid H_1 \text{ is true})$$

$$= P(3 \text{ or } 4 \text{ are black from box (2)})$$

$$= \frac{{}^5C_3 {}^{50}C_1}{{}^{100}C_4} + \frac{{}^5C_4 {}^{50}C_0}{{}^{100}C_4}$$

$$\approx 0.24992 + 0.05873$$

$$\approx 0.30865$$

$$\text{and } 1 - \beta \approx 0.691 \text{ or } 69.1\%$$

$\therefore$ the new decision rule does not give greater power.

6 H_0 : he rolls a 6, 9 times every 10 rolls $\therefore p = 0.9$

H_1 : $p < 0.9$

a i Let X = number of 6s in six rolls

$\therefore X \sim B(6, 0.9)$

$$\begin{aligned} & P(5 \text{ or } 6 \text{ sixes} \mid p = 0.9) \\ &= P(X \geq 5 \mid X \sim B(6, 0.9)) \\ &= 1 - P(X \leq 4 \mid X \sim B(6, 0.9)) \\ &\approx 0.886 \end{aligned}$$

$$\begin{aligned} \text{ii} \quad & P(1 \text{ six in 6 rolls} \mid X \sim B(6, \frac{1}{6})) \\ &= P(X \geq 5 \mid X \sim B(6, \frac{1}{6})) \\ &\approx P(X \leq 4 \mid X \sim B(6, \frac{1}{6})) \\ &\approx 0.000664 \end{aligned}$$

b H_0 : $p = 0.9$ against H_1 : $p < 0.9$

We will accept H_0 if $X \geq 4$ and
reject H_0 if $X \leq 3$.

$$\begin{aligned} \text{i} \quad & P(\text{accept } H_0 \mid X \sim B(6, \frac{1}{6})) \\ &= P(X \geq 4 \mid X \sim B(6, \frac{1}{6})) \\ &= 1 - P(X \leq 3 \mid X \sim B(6, \frac{1}{6})) \\ &\approx 0.00870 \end{aligned}$$

$$\begin{aligned} \text{ii} \quad & P(\text{reject } H_0 \mid X \sim B(6, 0.9)) \\ &= P(X \leq 3 \mid X \sim B(6, 0.9)) \\ &\approx 0.0159 \end{aligned}$$

Note: i gives a Type II error

ii gives a Type I error

7 $X \sim \text{Po}(m)$ H_0 : $m = 3$

H_1 : $m = 4$

$$\text{Let } S = \sum_{i=1}^9 x_i$$

The critical region is $S \leq 37$ for H_0 acceptance and
 $S \geq 38$ for H_1 acceptance.

$$\begin{aligned} \text{a} \quad & \alpha = P(\text{Type I error}) \\ &= P(\text{rejecting } H_0 \mid H_0 \text{ is true}) \\ &= P(S \geq 38 \mid m = 3) \\ &= 1 - P(S \leq 37 \mid S \sim \text{Po}(3 \times 9)) \\ &\approx 0.0263 \\ \therefore \alpha &\approx 2.63\% \end{aligned}$$

b Power

$$\begin{aligned} &= 1 - P(\text{Type II error}) \\ &= 1 - P(\text{accepting } H_0 \mid H_1 \text{ is true}) \\ &= 1 - P(S \leq 37 \mid m = 4) \\ &= 1 - P(S \leq 37 \mid S \sim \text{Po}(4 \times 9)) \\ &\approx 0.391 \quad (\text{or } \approx 39.1\%) \end{aligned}$$

8 $X \sim \text{Geo}(p)$. H_0 : $p = 0.25$ and H_1 : $p = 0.38$

a $S = \sum_{i=1}^{12} x_i$ is $\text{NB}(12, p)$ and the critical region for rejecting
 H_0 is $S \leq 29$ or $S \geq 72$

$$\begin{aligned} \text{b} \quad & P(\text{Type I error}) \\ &= P(\text{rejecting } H_0 \mid H_0 \text{ is true}) \\ &= P(S \leq 29 \text{ or } S \geq 72 \mid p = 0.25) \\ &= P(S \leq 29 \mid p = 0.25) + 1 - P(S \leq 71 \mid p = 0.25) \end{aligned}$$

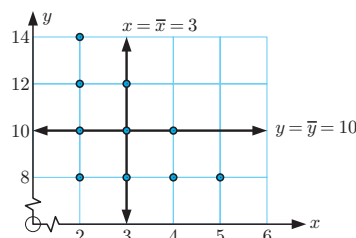
$$\begin{aligned} &= \sum_{i=12}^{29} \binom{i-1}{11} (0.25)^{12} (0.75)^{i-12} \\ &\quad + 1 - \sum_{i=12}^{71} \binom{i-1}{11} (0.25)^{12} (0.75)^{i-12} \\ &\approx 0.03903 + 1 - 0.96154 \\ &\approx 0.0775 \end{aligned}$$

c P(Type II error)

$$\begin{aligned} &= P(\text{accepting } H_0 \mid H_1 \text{ is true}) \\ &= P(30 \leq S \leq 71 \mid p = 0.38) \\ &= P(S \leq 71 \mid p = 0.38) - P(S \leq 29 \mid p = 0.38) \\ &\approx 0.99997 - 0.42149 \\ &\approx 0.578 \end{aligned}$$

EXERCISE 1.1

1 a



b There is weak to moderate negative correlation.

c $\bar{x} = 3$, $\bar{y} = 10$ are drawn in a.

These lines and the data points within the four quadrants do support the decision in b, as there are 4 data points in the 'negative' quadrants and only 1 in the 'positive' quadrants.

d From technology, $\sigma_x = 1$ and $\sigma_y = 2$

$$\begin{aligned} r &= \frac{1}{n} \sum_{i=1}^n (z_{x_i})(z_{y_i}) \\ &= \frac{1}{10} \sum_{i=1}^{10} \left(\frac{x_i - 3}{1} \right) \left(\frac{y_i - 10}{2} \right) \\ &= \frac{1}{20} \sum_{i=1}^{10} (x_i - 3)(y_i - 10) \\ &= \frac{1}{20} [(-1 \times -2) + (-1 \times 0) + (-1 \times 2) + (-1 \times 4) \\ &\quad + (0 \times -2) + (0 \times 0) + (0 \times 2) + (1 \times -2) \\ &\quad + (1 \times 0) + (2 \times -2)] \\ &= \frac{1}{20} [2 - 2 - 4 - 2 - 4] \\ &= \frac{1}{20} [-10] \\ &= -0.5 \end{aligned}$$

$$2 \text{ a} \quad \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$\begin{aligned} &= \sum_{i=1}^n (x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x} \bar{y}) \\ &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - \bar{y}(n \bar{x}) - \bar{x}(n \bar{y}) + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \end{aligned}$$

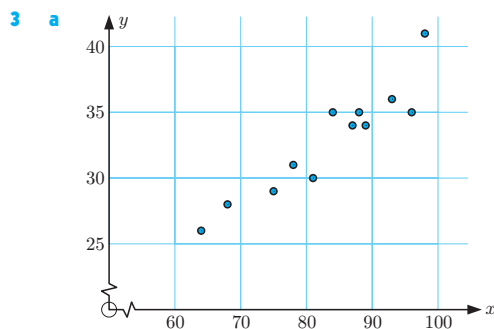
$$\begin{aligned}
 \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
 &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \\
 &= \sum_{i=1}^n x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\
 &= \sum_{i=1}^n x_i^2 - n\bar{x}^2
 \end{aligned}$$

$$\text{Likewise } \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

$$\text{Thus } r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}}$$

$$\begin{aligned}
 \text{b } \sum x_i y_i &= 16 + 20 + 24 + 28 + 24 + 30 + 36 + 32 \\
 &\quad + 40 + 40 \\
 &= 290 \\
 \sum x_i^2 &= 4 + 4 + 4 + 4 + 9 + 9 + 9 + 16 + 16 + 25 \\
 &= 4 \times 4 + 3 \times 9 + 2 \times 16 + 25 \\
 &= 100 \\
 \sum y_i^2 &= 8^2 + 10^2 + 12^2 + 14^2 + 8^2 + 10^2 + 12^2 + 8^2 \\
 &\quad + 10^2 + 8^2 \\
 &= 4 \times 8^2 + 3 \times 10^2 + 2 \times 12^2 + 14^2 \\
 &= 1040 \\
 \therefore r &= \frac{290 - 10 \times 3 \times 10}{\sqrt{(100 - 10 \times 3^2)(1040 - 10 \times 10^2)}} \\
 &= \frac{-10}{\sqrt{10 \times 40}} \\
 &= \frac{-10}{20} \\
 &= -0.5
 \end{aligned}$$

Technology gives the values for $\bar{x}$, $\bar{y}$, n , $\sum x_i^2$, $\sum y_i^2$, $\sum x_i y_i$, and r .



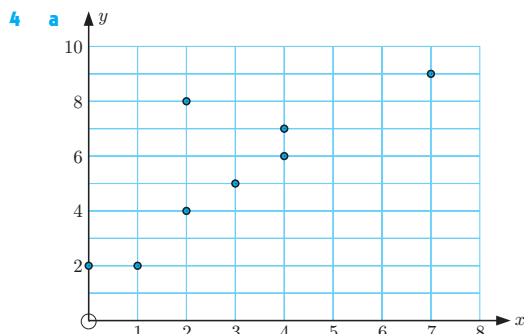
b The data appears to have a strong positive correlation.

c From technology, $\sum xy = 33\,321$

$$\bar{x} = 83.42, \quad \sum x = 1001, \quad \sum x^2 = 84\,749$$

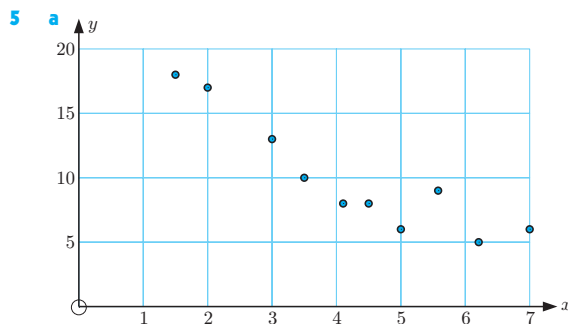
$$\bar{y} = 32.83, \quad \sum y = 394, \quad \sum y^2 = 13\,126$$

$$\begin{aligned}
 \therefore r &= \frac{\sum_{i=1}^{12} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \sqrt{\sum_{i=1}^n y_i^2 - n\bar{y}^2}} \\
 &= \frac{33\,321 - 12(83.42)(32.83)}{\sqrt{84\,749 - 12 \times 83.42^2} \sqrt{13\,126 - 12 \times 32.83^2}} \\
 &\approx 0.935 \quad \text{confirming a strong positive correlation} \\
 &\quad \text{(Technology could have been used here.)}
 \end{aligned}$$



b Using technology, $r \approx 0.816$

c There appears to be a moderate positive correlation.



b Negative, as a line of best fit by eye would have negative slope.

c From technology, $r \approx -0.911$

d There appears to be a strong negative correlation between the variables.

e As the distance from goal increases, the number of successful shots generally decreases.

f Yes, as the distance from the goal does affect the player's ability to become successful.

6 a For

x	1	2	3
y	1	3	4

 $r \approx 0.982$

b i For $u = 2x + 1$, $v = 3y - 1$ ii $r \approx 0.982$

u	3	5	7
v	2	8	11

c i For $u = -2x + 1$, $v = 3y - 1$ ii $r \approx -0.982$

u	-1	-3	-5
v	2	8	11

d i For $u = 2x + 1$, $v = -3y - 1$ ii $r \approx -0.982$

u	3	5	7
v	-4	-10	-13

- e i** For $u = -2x + 1$, $v = -3y - 1$ **ii** $r \approx 0.982$

u	-1	-3	-5
v	-4	-10	-13

- f** For $u = ax + b$ and $v = cx + d$, u and v have the same correlation coefficient, r , as x and y if $ac > 0$, or have correlation coefficient $-r$ if $ac < 0$.

EXERCISE 1.2

1 $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$

a $\text{Cov}(X, X) = E(X^2) - [E(X)]^2$
 $= \text{Var}(X)$

b $\text{Cov}(X, X + Y)$
 $= E(X(X + Y)) - E(X)E(X + Y)$
 $= E(X^2 + XY) - E(X)E(X + Y)$
 $= E(X^2) + E(XY) - E(X)[E(X) + E(Y)]$
 $= E(X^2) + E(XY) - [E(X)]^2 - E(X)E(Y)$
 $= E(X^2) - [E(X)]^2 + E(XY) - E(X)E(Y)$
 $= \text{Var}(X) + \text{Cov}(X, Y)$
 $= \text{Cov}(X, X) + \text{Cov}(X, Y) \quad \{\text{from a}\}$

c If $X = c$, $\text{Cov}(XY) = E(cY) - E(c)E(Y)$
 $= cE(Y) - cE(Y)$
 $= 0 \quad \text{for any random variable } Y$

2 $\text{Cov}(X + Y, X - Y)$
 $= E((X + Y)(X - Y)) - E(X + Y)E(X - Y)$
 $= E(X^2 - Y^2) - [E(X) + E(Y)][E(X) - E(Y)]$
 $= E(X^2) - E(Y^2) - [E(X)]^2 + [E(Y)]^2$
 $= E(X^2) - [E(X)]^2 - \{E(Y^2) - [E(Y)]^2\}$
 $= \text{Var}(X) - \text{Var}(Y)$

or $\text{Cov}(X, X) - \text{Cov}(Y, Y)$

3 $\text{Var}(X + Y)$
 $= E((X + Y)^2) - [E(X + Y)]^2$
 $= E(X^2 + 2XY + Y^2) - [E(X) + E(Y)]^2$
 $= E(X^2) + 2E(XY) + E(Y^2)$
 $\quad - [E(X)]^2 - 2E(X)E(Y) - [E(Y)]^2$
 $= E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2$
 $\quad + 2(E(XY) - E(X)E(Y))$
 $= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

4 $\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$
 $\therefore \rho^2 = \frac{\{\text{Cov}(X, mX + c)\}^2}{\text{Var}(X)\text{Var}(mX + c)}$
 $\therefore \rho^2 = \frac{\{E(mX^2 + cX) - E(X)E(mX + c)\}^2}{\text{Var}(X)\{m^2\text{Var}(X)\}}$
 $\therefore \rho^2 = \frac{\{mE(X^2) + cE(X) - m[E(X)]^2 - cE(X)\}^2}{m^2[\text{Var}(X)]^2}$
 $\therefore \rho^2 = \frac{\{m(E(X^2) - [E(X)]^2)\}^2}{m^2[\text{Var}(X)]^2}$
 $\therefore \rho^2 = \frac{\{m\text{Var}(X)\}^2}{m^2[\text{Var}(X)]^2}$
 $\therefore \rho^2 = 1 \quad \{\text{as } m \neq 0\}$
 $\therefore \rho = \pm 1$

- 5** The product moment correlation coefficient of U and V

$$\begin{aligned} &= \frac{\text{Cov}(U, V)}{\sqrt{\text{Var}(U)\text{Var}(V)}} \\ &= \frac{\text{Cov}(a + bX, c + dY)}{\sqrt{\text{Var}(a + bX)\text{Var}(c + dY)}} \\ &= \frac{E((a + bX)(c + dY)) - E(a + bX)E(c + dY)}{\sqrt{b^2\text{Var}(X)d^2\text{Var}(Y)}} \\ &= \frac{E(ac + adY + bcX + bdXY) - [a + bE(X)][c + dE(Y)]}{|bd|\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \frac{ac + adE(Y) + bcE(X) + bdE(XY) - ac - adE(Y) - bcE(X) - bdE(X)E(Y)}{|bd|\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \frac{bd[E(XY) - E(X)E(Y)]}{|bd|\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \pm 1 \times \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \pm \rho \end{aligned}$$

6 $\sigma_X^2 = 1$, $\sigma_Y^2 = 9$, $\rho = \frac{1}{9}$

From **3**,

$$\begin{aligned} \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) \\ &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y \\ &\quad \left\{ \text{as } \rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \right\} \end{aligned}$$

$$\begin{aligned} \therefore \text{Var}(X + Y) &= 1 + 9 + 2\left(\frac{1}{9}\right) \times 1 \times 3 \\ &= 10\frac{2}{3} \end{aligned}$$

Now the correlation coefficient between X and $X + Y$

$$\begin{aligned} &= \frac{\text{Cov}(X, X + Y)}{\sqrt{\text{Var}(X)\text{Var}(X + Y)}} \\ &= \frac{\text{Var}(X) + \text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(X + Y)}} \\ &= \frac{\sigma_X + \rho\sigma_X\sigma_Y}{\sigma_X \times \sqrt{\frac{32}{3}}} \\ &= \frac{1 + \frac{1}{9}(1)(3)}{1 \times \sqrt{\frac{32}{3}}} \\ &= \frac{4}{3} \times \sqrt{\frac{3}{32}} \\ &= \frac{1}{\sqrt{6}} \end{aligned}$$

EXERCISE 1.3

- 1 a** From technology, $r \approx 0.67169$
 The regression of Y on X is $y \approx 0.88278x + 5.5957$
 The regression of X on Y is $x \approx 0.51108y + 2.7382$
- b** If $y = 11$, $x \approx 0.51108 \times 11 + 2.7382$
 ≈ 8.36
 is the estimated score for Mathematics.
- c** If $x = 18$, $y \approx 0.88278 \times 18 + 5.5957$
 ≈ 21.5
 is the estimated score for Physics.

- d** $y = 11$ is within the given y -values whereas $x = 18$ is outside the given x -values. As we cannot guarantee that the linear model continues outside the given values, we expect the estimated score for Mathematics to be more accurate.
- 2 a** From technology, $r \approx 0.90584$
 The regression line for Y on X is $y \approx 16.960x - 35.528$
 The regression line for X on Y is $x \approx 0.048380y + 2.7031$
- b** If $y = 60$, $x = 0.048380 \times 60 + 2.7031$
 ≈ 5.61 is the estimated cholesterol level.
- c** If $x = 5.8$, $y = 16.960 \times 5.8 - 35.528$
 ≈ 63 is the estimated resting heart rate.
- 3 a** From technology, $r \approx -0.94046$
 The regression line for Y on X is $y \approx -1.2768x + 218.70$
 The regression line for X on Y is $x \approx -0.69269y + 159.72$
- b** When $x = 65$, $y \approx -1.2768 \times 65 + 218.70$
 ≈ 135.7 sec
 is the estimated time to swim 200 m breaststroke.
- 4 a** **i** $r \approx 0.78400$ (moderate, positive)
ii $r \approx 0.94774$ (strong, positive)
iii $r \approx 0.05472$ (very weak, positive)

b $w\% = \frac{y}{x} \times 100\%$

Men

$Wt(x)$	89	88	66	59	93	73	82	77	100	67
$\%BF(w)$	13.5	15.9	13.6	16.9	23.7	17.8	15.9	14.3	19.0	17.9

Women

$Wt(y)$	57	68	69	59	62	59	56	66	72
$\%BF(w)$	29.8	32.4	34.8	30.5	29.0	25.4	28.6	33.3	33.3

- c** **i** $r \approx 0.34181$ (weak, positive)
ii $r \approx 0.79075$ (moderate, positive)
iii $r \approx -0.4620$ (weak, negative)

EXERCISE 1.4

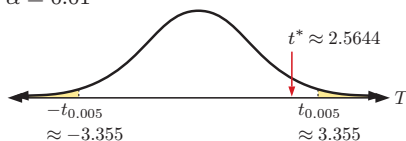
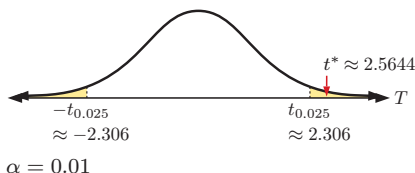
- 1** $H_0: \rho = 0$ against $H_1: \rho \neq 0$
 $n = 10$, $\therefore \nu = df = 8$
 (1) $r \approx 0.67169$

(2) $t^* \approx 0.67169 \times \sqrt{\frac{8}{1 - 0.67169^2}}$
 $\therefore t^* \approx 2.5644$

(3) **Either** $p\text{-value} = 2 \times P(T < 2.5644)$
 ≈ 0.0334

which is < 0.05 but > 0.01

or $\alpha = 0.05$



$\therefore t^*$ is inside the critical region for $\alpha = 0.05$ but outside it for $\alpha = 0.01$.

- (4) At a 5% level, we accept H_0 that the data is not correlated. At a 1% level, we do not accept H_0 . We accept that the data is correlated.

- 2** $H_0: \rho = 0$ against $H_1: \rho \neq 0$
 $n = 10$, $\therefore \nu = df = 8$

(1) $r \approx 0.90584$

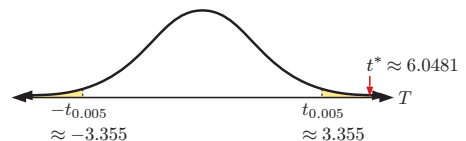
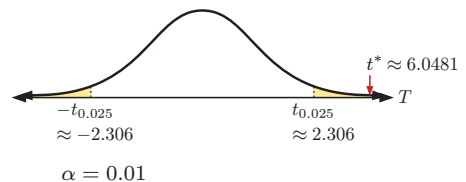
(2) $t^* \approx 0.90584 \times \sqrt{\frac{8}{1 - 0.90584^2}}$

$\therefore t^* \approx 6.0481$

(3) **Either** $p\text{-value} = 2 \times P(T > 6.0481)$
 ≈ 0.000307

which is < 0.05 and is < 0.01

or $\alpha = 0.05$



and for both levels t^* does not lie in the critical region.

- (4) At both the 5% level and 1% level we accept H_0 , that the data is not correlated.

- 3** $H_0: \rho = 0$ against $H_1: \rho \neq 0$
 (1) $r \approx 0.69558$

(2) $t^* \approx 0.69558 \times \sqrt{\frac{7}{1 - 0.69558^2}}$
 $\therefore t^* \approx 2.5615$

(3) $p\text{-value} = 2 \times P(T > 2.5615)$
 ≈ 0.0375

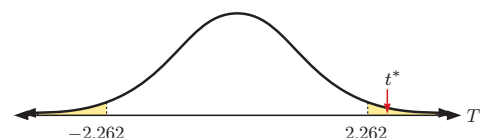
which is < 0.05 but > 0.01

- (4) At a 5% level we accept H_0 that the data is not correlated. At a 1% level we do not accept H_0 . We accept that the data is correlated.

- 4** $H_0: \rho = 0$ against $H_1: \rho \neq 0$
 (1) $r \approx 0.606886$

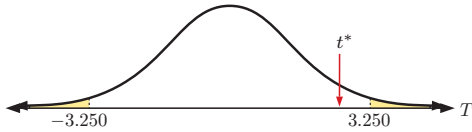
(2) $t^* \approx 0.606886 \times \sqrt{\frac{9}{1 - 0.606886^2}}$
 $\therefore t^* \approx 2.291$

(3) **For** $\alpha = 0.05$, $t_{\frac{\alpha}{2}} = t_{0.025} \approx 2.262$



t^* lies in the critical region.

For $\alpha = 0.01$, $t_{\frac{\alpha}{2}} = t_{0.005} \approx 3.250$



t^* does not lie in the critical region.

- (4) So, at a 5% level we do not accept the data is uncorrelated but at a 1% level we do.

At a 1% level we accept that $\rho = 0$ and therefore that X and Y are independent. However, we do not accept this at a 5% level.

$$5 \quad t^* = r \sqrt{\frac{n-2}{1-r^2}} = -0.5 \sqrt{\frac{14}{1-0.25}} \approx -2.16$$

$$p\text{-value} = 2 \times P(T > 2.16) \\ \approx 0.0485$$

which is less than 0.05 but not less than 0.01.

So, at a 5% level we reject $H_0: \rho = 0$ in favour of

$H_1: \rho \neq 0$, and at a 1% level we do not reject H_0 .

Thus, at a 1% level we accept that $\rho = 0$ and therefore X and Y are uncorrelated.

However, we do not accept this at a 5% level.

$$6 \quad \text{If } r = 1, \text{ the } t^* \text{ calculation would be } t^* = 1 \times \sqrt{\frac{n}{1-1}}$$

which is undefined.

Thus the hypothesis test is unusable.

The test is unnecessary as perfect positive correlation exists.

- 7 We test $H_0: \rho = 0$ against $H_1: \rho \neq 0$ with $r = 0.6$

$$\text{The } t\text{-statistic is } t^* = 0.6 \sqrt{\frac{n-2}{1-0.6^2}} \\ = 0.75 \sqrt{n-2}$$

For infinitely many degrees of freedom the t -distribution approximates the normal distribution $Z \sim N(0, 1)$ where

$$t_{\frac{\alpha}{2}} \approx z_{\frac{\alpha}{2}} = z_{0.025} \approx 1.960$$

If $t^* > 1.960$ then $0.75 \sqrt{n-2} > 1.960$

$$\therefore \sqrt{n-2} > 2.613 \dots$$

$$\therefore n-2 > 6.829 \dots$$

$$\therefore n > 8.829 \dots$$

$$\therefore n \geq 9$$

If $n = 9$, $\nu = 7$, $t_{0.025} \approx 2.365$, $t^* \approx 1.984$

$$\therefore t^* < t_{0.025}$$

If $n = 10$, $\nu = 8$, $t_{0.025} \approx 2.306$, $t^* \approx 2.121$

$$\therefore t^* < t_{0.025}$$

If $n = 11$, $\nu = 9$, $t_{0.025} \approx 2.262$, $t^* \approx 2.250$

$$\therefore t^* < t_{0.025}$$

If $n = 12$, $\nu = 10$, $t_{0.025} \approx 2.228$, $t^* \approx 2.372$

$$\therefore t^* > t_{0.025}$$

For $n = 12$ we would reject $H_0: \rho = 0$ in favour of

$H_1: \rho \neq 0$

$\therefore$ a sample size of at least 12 is required to conclude, at a 1% level of significance, that X and Y are correlated.

- 8 We test $H_0: \rho = 0$ against $H_1: \rho \neq 0$ with $n = 20$, $\nu = 18$

$$\text{The } t\text{-statistic is } t^* = r \sqrt{\frac{18}{1-r^2}}$$

$$\text{For } \alpha = 0.05, t_{\frac{\alpha}{2}} = t_{0.025} \approx 2.1009$$

We reject $H_0: \rho = 0$ if $|t^*| > 2.1009$

$$\therefore |r| \sqrt{\frac{18}{1-r^2}} > 2.1009$$

$$\therefore \frac{r^2 \times 18}{1-r^2} > 4.4139$$

$$\therefore 18r^2 > 4.4139 - 4.4139r^2$$

$$\therefore 22.4139r^2 > 4.4139$$

$$\therefore r^2 > \frac{4.4139}{22.4139} = 0.196926$$

$$\therefore \sqrt{r^2} > \sqrt{0.196926} \approx 0.44376$$

$$\therefore |r| > 0.44376$$

$\therefore$ least value of $|r|$ is 0.444 (to 3 s.f.)

REVIEW SET A

- 1 a i $E(X_1 + 2X_2 + 3X_3)$
 $= E(X_1) + 2E(X_2) + 3E(X_3)$
 $= \mu + 2\mu + 3\mu$
 $= 6\mu$
 $\text{Var}(X_1 + 2X_2 + 3X_3)$
 $= \text{Var}(X_1) + 2^2 \text{Var}(X_2) + 3^2 \text{Var}(X_3)$
 $= \sigma^2 + 4\sigma^2 + 9\sigma^2$
 $= 14\sigma^2$
 $\therefore X_1 + 2X_2 + 3X_3$ has mean 6μ and a standard deviation $\sigma\sqrt{14}$.
- ii $E(2X_1 - 3X_2 + X_3)$
 $= 2E(X_1) - 3E(X_2) + E(X_3)$
 $= 2\mu - 3\mu + \mu$
 $= 0$
 $\text{Var}(2X_1 - 3X_2 + X_3)$
 $= 2^2 \text{Var}(X_1) + (-3)^2 \text{Var}(X_2) + \text{Var}(X_3)$
 $= 4\sigma^2 + 9\sigma^2 + \sigma^2$
 $= 14\sigma^2$
 $\therefore 2X_1 - 3X_2 + X_3$ has mean 0 and standard deviation $\sigma\sqrt{14}$.
- b We use $E(X^2) = \text{Var}(X) + [E(X)]^2$
 $E([X_1 - X_2]^2)$
 $= E(X_1^2 - 2X_1X_2 + X_2^2)$
 $= E(X_1^2) - 2E(X_1X_2) + E(X_2^2)$
 $= \sigma^2 + \mu^2 - 2E(X_1)E(X_2) + \sigma^2 + \mu^2$
 $= 2\sigma^2 + 2\mu^2 - 2\mu\mu$
 $= 2\sigma^2$
- 2 a Let X be the number of buses needed to get a correct one.
 $X \sim \text{Geo}(0.35)$
 i $P(X \leq 4) \approx 0.821$
 ii $E(X) = \frac{1}{p} = \frac{1}{0.35} \approx 2.86$ (about 3)

- b** Let Y be the bus that the student will take to school.
 $Y \sim \text{NB}(3, 0.35)$

i $P(Y = 7) = \binom{6}{2} (0.35)^3 (0.65)^4 \approx 0.115$

ii $E(Y) = \frac{r}{p} = \frac{3}{0.35} \approx 8.57$

$\therefore$ the average number of buses until the correct one arrives is approximately 9 buses.

iii $P(Y \leq 5)$
 $= P(Y = 3 \text{ or } 4 \text{ or } 5)$
 $= \binom{2}{2} (0.35)^3 + \binom{3}{2} (0.35)^3 (0.65)$
 $+ \binom{4}{2} (0.35)^3 (0.65)^2$
 $= (0.35)^3 [1 + 3 \times 0.65 + 6 \times (0.65)^2]$
 ≈ 0.235

3 a As $\sum p_i = 1$, $5c = 1$
 $\therefore c = 0.2$

b $\mu_X = \sum p_i x_i$
 $= -3(0.2) - 1(0.2) + 1(0.2) + 3(0.2) + 5(0.2)$
 $= 5(0.2)$
 $= 1$

c $P(X > 1) = P(X = 3 \text{ or } 5)$
 $= 0.2 + 0.2$
 $= 0.4$

d $\text{Var}(X) = \sum p_i x_i^2 - (\mu_X)^2$
 $= 0.2[3^2 + 1^2 + 1^2 + 3^2 + 5^2] - 1^2$
 $= 0.2 \times 45 - 1$
 $= 8$

- 4** Let X be the number who prefer right leg kick.
 $X \sim \text{B}(n, 0.75)$

a i $n = 20$
 $P(X = 14) \approx 0.169$

ii $n = 20$
 $P(X \geq 15) = 1 - P(X \leq 14)$
 ≈ 0.617

- b** $X \sim \text{B}(1050, 0.75)$
 Now $np > 10$ and $n(1-p) > 10$
 $\therefore$ we can approximate X by a normal variate with

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{np(1-p)}$$

$$= 787.5 \quad \quad \quad = 14.03$$

i $P(70\% \text{ prefer right leg})$
 $= P(X = 0.7 \times 1050)$
 $= P(X = 735)$
 $\approx P(734.5 < X^* < 735.5)$
 $\approx 0.000\,026\,0$

ii $P(\text{no more than } 25\% \text{ prefer left})$
 $= P(X \geq 787.5)$
 $= 1 - P(X \leq 787)$
 ≈ 0.514

- 5 a** If $X \sim \text{B}(n, p)$, then X has PDF

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $\sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = [p + (1-p)]^n$

$$\therefore G(t) = \sum_{x=0}^n t^x \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^n \binom{n}{x} (pt)^x (1-p)^{n-x}$$

$$= (pt + (1-p))^n$$

$$= (1-p+pt)^n \quad \text{for } t = 0, 1, 2, 3, \dots, n$$

b i $G'(t) = n(1-p+pt)^{n-1} \times p$
 $= np(1-p+pt)^{n-1}$

$\therefore G'(1) = np(1)^{n-1} = np$

$\therefore E(X) = np$

ii $G''(t) = n(n-1)p(1-p+pt)^{n-2} \times p$

$\therefore G''(1) = n(n-1)p^2$

Thus, $\text{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$
 $= n(n-1)p^2 + np - n^2p^2$
 $= np[(n-1)p + 1 - np]$
 $= np[1-p]$
 $= np(1-p)$

- 6** Let X denote the number of hours lost to sickness.

$$X \sim \text{N}(\mu, 67^2)$$

$\therefore \bar{X} \sim \text{N}(\mu, \frac{67^2}{375})$ {CL Theorem}

We need to find $P(|\bar{X} - \mu| > 10)$

$$= P(\bar{X} - \mu > 10 \text{ or } \bar{X} - \mu < -10)$$

$$= 1 - P(-10 < \bar{X} - \mu < 10)$$

$$= 1 - P\left(\frac{-10}{\frac{67}{\sqrt{375}}} < \frac{\bar{X} - \mu}{\frac{67}{\sqrt{375}}} < \frac{10}{\frac{67}{\sqrt{375}}}\right)$$

$$= 1 - P(-2.8903 < Z < 2.8903)$$

$$\approx 0.003\,85$$

- 7 a** Using technology, $\bar{x} = 12.275$

$$s_n \approx 1.5164 \quad (s_{n-1} \approx 1.5357)$$

- b** σ is unknown, so we use s_{n-1} as an unbiased estimate of it.

If X is the number of points held by an employee,

$$X \sim t(39).$$

95% CI for μ is

$$\bar{x} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{where } t_{\frac{\alpha}{2}} \approx 1.685$$

$$\text{Using technology, } 11.8 < \mu < 12.8$$

- 8 a** As we have the same group of students, observations are not independent.

So, the difference between means is not appropriate.

b

Student	A	B	C	D	E	F	G	H	I	J
Pre-test (X_1)	12	13	11	14	10	16	14	13	13	12
Post-test (X_2)	11	14	16	13	12	18	15	14	15	11
Difference (D)	-1	1	5	-1	2	2	1	1	2	-1

$$D = X_2 - X_1$$

$$\mu_D = 1.1, \quad s_{n-1} \approx 1.8529 \quad \{\text{using technology}\}$$

$s_{n-1} \approx 1.8529$ is an unbiased estimate of σ .

$D \sim t(9)$ and the 90% confidence interval for D is
 $0.025\,89 < D < 2.1741$

- c $H_0: \mu_D = 0$ (there is no improvement)
 $H_1: \mu_D > 0$ (there is an improvement)
 We perform a 1-tailed test at a 5% level with 9 df and $D \sim t(9)$.
 $t^* \approx 1.8773$ and $p\text{-value} \approx 0.0466$.
 Since the $p\text{-value} < 0.05$, we reject H_0 in favour of H_1 .
 $\therefore$ at a 5% level of significance, there has been an improvement.

- 9 (1) $H_0: \mu = 68$ (unchanged)
 $H_1: \mu < 68$ (decreased)
 (2) As σ is unknown, we use s_{n-1} as an unbiased estimate of σ .

$$s_{n-1} = \sqrt{\frac{n}{n-1}} s_n = \sqrt{\frac{42}{41}} \times 1.732$$

$$\therefore s_{n-1} \approx 1.753$$
 The null distribution is $t(41)$.
 (3) The test statistic is t^* where

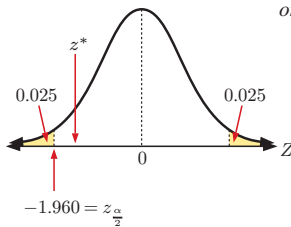
$$t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} = \frac{65 - 68}{\frac{1.753}{\sqrt{42}}} \approx -11.09$$
 (4) We reject H_0 if the $p\text{-value} < 0.05$.
 (5) $p\text{-value} = 0$
 (6) $\therefore$ we reject H_0 in favour of H_1 .
 (7) There is sufficient evidence at a 5% level to accept H_1 that the diet and exercise program has reduced Yarni's resting pulse rate.

- 10 (1) $H_0: \mu = 7.82$ (no change)
 $H_1: \mu \neq 7.82$
 (2) As σ is known ($\sigma = 1.83$) we use the Z -distribution (the null distribution).

(3) The test statistic is $z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

$$\therefore z^* = \frac{7.55 - 7.82}{\frac{1.83}{\sqrt{48}}} \approx -1.022$$

- (4) We reject H_0 if:
 z^* is in the critical region or $p\text{-value} < 0.05$
 (5) or $p\text{-value} \approx 2 \times P(Z \leq -1.022) \approx 0.3067$



- (6) z^* does not lie in the critical region. or The $p\text{-value}$ is not less than 0.05.
 $\therefore$ we do not reject H_0 . $\therefore$ we do not reject H_0 .
 (7) There is insufficient evidence to reject H_0 at a 5% level of significance. So, we accept that there has been no change in the colony's mean weight at this level.

- 11 a By definition, $\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$
 b If $\rho = 0$, then $\text{Cov}(X, Y) = 0$ {from a}
 $\therefore E(XY) - E(X)E(Y) = 0$
 $\{ \text{as } \text{Cov}(X, Y) = E(XY) - E(X)E(Y) \}$
 $\therefore E(XY) = E(X)E(Y)$
 $\therefore X$ and Y are independent random variables.
 c i Using technology, $r \approx 0.49594$
 ii $H_0: \rho = 0$ against $H_1: \rho \neq 0$
 The test statistic is

$$t^* = r \sqrt{\frac{n-2}{1-r^2}} = 0.49594 \sqrt{\frac{9}{1-0.49594^2}}$$

$$\therefore t^* \approx 1.713 \text{ with } \nu = n-2 = 9$$

$$p\text{-value} = 2 \times P(T > |t^*|)$$

$$= 2 \times P(T > 1.713)$$

$$\approx 0.1209 \text{ which is not } < 0.05$$
 Thus, at a 5% level, we do not reject H_0 in favour of H_1 .
 That is, we accept $\rho = 0$, in which case X and Y are independent random variables.

12 a $\text{Var}(\bar{X})$ b $E(\bar{X})$

$$= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right)$$

$$= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right)$$

$$= \frac{1}{n^2} \times n\sigma^2 = \frac{1}{n} \sum_{i=1}^n E(X_i)$$

$$= \frac{\sigma^2}{n} = \frac{1}{n} \times n\mu$$

$$= \mu$$

c $E(S_n^2)$

$$= E\left(\frac{1}{n} \left\{ \left(\sum_{i=1}^n X_i^2 \right) - n\bar{X}^2 \right\}\right)$$

$$= \frac{1}{n} \left\{ E\left(\sum_{i=1}^n X_i^2\right) - nE(\bar{X}^2) \right\}$$

$$= \frac{1}{n} \left\{ \sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right\}$$

$$= \frac{1}{n} \left\{ \sum_{i=1}^n (\text{Var}(X_i) + [E(X_i)]^2) - n(\text{Var}(\bar{X}) + [E(\bar{X})]^2) \right\}$$

{since $\text{Var}(U) = E(U^2) - [E(U)]^2$ for any random variable U }

$$= \frac{1}{n} \left\{ \sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \right\}$$

$$= \frac{1}{n} \{ n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2 \}$$

$$= \sigma^2 + \cancel{\mu^2} - \frac{\sigma^2}{n} - \cancel{\mu^2}$$

$$= \sigma^2 \left(1 - \frac{1}{n}\right)$$

$$= \left(\frac{n-1}{n} \right) \sigma^2$$

$$\neq \sigma^2$$

$\therefore S_{n-1}^2$ is a biased estimator of σ^2 .

$$\begin{aligned} \text{d } E(S_{n-1}^2) &= \left(\frac{n}{n-1} \right) E(S_n^2) \\ &= \frac{n}{n-1} \times \frac{n-1}{n} \sigma^2 \\ &= \sigma^2 \end{aligned}$$

$\therefore S_{n-1}^2$ is an unbiased estimator of σ^2 .

REVIEW SET B

1 a i Since $\sum p_i = 1$,
 $0.3 + 0.2 + 0.2 + P(X = 6) = 1$
 $\therefore P(X = 6) = 0.3$

ii $E(X) = \sum p_i x_i$
 $= -5(0.3) - 1(0.2) + 3(0.2) + 6(0.3)$
 $= 0.7$

$\therefore$ expected return is 70 cents/game.

iii At 50 cents/game, they would on average lose 20 cents each game. However, if they charge \$1/game, they would expect to gain 30 cents each game.

b i $E(Y) = -3(0.5) + 2(0.3) + 5(0.2)$
 $= 0.1$

$\therefore$ expected return is 10 cents/game.

ii

game 2	$X + Y$ values			
(0.2) 5	0	4	8	11
(0.3) 2	-3	1	5	8
(0.5) -3	-8	-4	0	3
	-5	-1	3	6
	(0.3)	(0.2)	(0.2)	(0.3)

game 1

$X + Y$	-8	-4	-3	0	1
prob.	0.15	0.10	0.09	0.16	0.06

$X + Y$	3	4	5	8	11
prob.	0.15	0.04	0.06	0.13	0.06

For example:

$$\begin{aligned} P(X + Y = -8) &= P(X = -5, Y = -3) \\ &= 0.3 \times 0.5 \\ &= 0.15 \end{aligned}$$

$$\begin{aligned} P(X + Y = -4) &= P(X = -1, Y = -3) \\ &= 0.2 \times 0.5 \\ &= 0.1, \text{ and so on} \end{aligned}$$

$$\begin{aligned} E(X + Y) &= -8(0.15) - 4(0.10) - 3(0.09) + \dots + 11(0.06) \\ &= 0.8 \end{aligned}$$

or $E(X + Y) = E(X) + E(Y)$
 $= 0.7 + 0.1$
 $= 0.8$

$\therefore$ expected return = 80 cents/game

iii Organiser's gain
 $= 1 \times (500 + 500 + 1000)$
 $- 0.7 \times 500 - 0.1 \times 500 - 0.8 \times 1000$
 $= \$800$

2 $X \sim B(1000, \frac{3}{5})$

a i $E(X) = np$
 $= 1000 \times \frac{3}{5}$
 $= 600$

ii $\text{Var}(X) = np(1-p)$
 $= 1000 \times \frac{3}{5} \times \frac{2}{5}$
 $= 240$
 $\therefore \sigma_X = \sqrt{240} \approx 15.5$

b $P(580 \leq X \leq 615)$
 $= P(X \leq 615) - P(X \leq 579)$
 $= 0.84147 - 0.09312$ {technology}
 ≈ 0.7484

c $P(579.5 \leq X \leq 615.5) \approx 0.7486$ {technology}

d If $X \sim B(n, p)$ where $np \geq 10$ and $n(1-p) \geq 10$ then $X \sim N(np, np(1-p))$
 and $P(a \leq X_1 \leq b)$ where X_1 is binomial
 $\approx P(a - \frac{1}{2} \leq X_2 \leq b + \frac{1}{2})$ where X_2 is normal
 {"the continuity correction"}

3 If $X \sim NB(r, p)$ then

$$\begin{aligned} P(X = x) &= \binom{x-1}{r-1} p^r (1-p)^{x-r} \\ \therefore P(X = k + r - 1) &= \binom{k+r-1-1}{r-1} p^r (1-p)^{k+r-1-r} \\ &\text{for } k = 1, 2, 3, 4, 5, \dots \\ &= \binom{k+r-2}{r-1} p^r (1-p)^{k-1} \end{aligned}$$

Letting $r = 1$, gives

$$\begin{aligned} P(X = k) &= \binom{k-1}{0} p (1-p)^{k-1} \\ &= (1-p)^{k-1} p \\ &\text{for } k = 1, 2, 3, 4, 5, \dots \end{aligned}$$

But for $X \sim \text{Geo}(p)$,

$$\begin{aligned} P(X = k) &= p(1-p)^{k-1} \text{ for } k = 1, 2, 3, 4, \dots \\ \therefore NB(1, p) &= \text{Geo}(p) \end{aligned}$$

4 a If $X \sim \text{Geo}(p)$, then X has PDF

$$P(X = x) = p(1-p)^{x-1} \text{ for } x = 1, 2, 3, 4, \dots$$

$$\begin{aligned} \therefore G(t) &= \sum_{x=1}^{\infty} t^x p(1-p)^{x-1} \\ &= \frac{p}{1-p} \sum_{x=1}^{\infty} [t(1-p)]^x \\ &= \frac{p}{1-p} \times \frac{u_1}{1-r} \text{ provided } |r| < 1 \\ &= \frac{p}{1-p} \times \frac{t(1-p)}{1-t(1-p)} \text{ provided } |t(1-p)| < 1 \\ &= \frac{pt}{1-t(1-p)} \text{ provided } |t| < \frac{1}{1-p} \\ G'(t) &= \frac{p[1-t(1-p)] - pt(-1+p)}{[1-t(1-p)]^2} \\ &= \frac{p}{(1-t+pt)^2} \end{aligned}$$

b $E(X) = G'(1)$

$$\begin{aligned} &= \frac{p}{[1-1+p]^2} \\ &= \frac{p}{p^2} \\ &= \frac{1}{p} \end{aligned}$$

$$G'(t) = p(1 - t + pt)^{-2}$$

$$G''(t) = -2p(1 - t + pt)^{-3} \times (-1 + p)$$

$$= -2p(1 - t + pt)^{-3}(p - 1)$$

$$\therefore G''(1) = -2p(p)^{-3}(p - 1)$$

$$= \frac{2(1 - p)}{p^2}$$

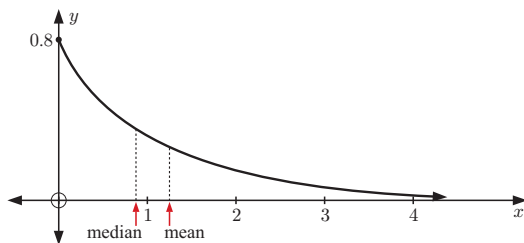
$$\text{Now } \text{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$$

$$= \frac{2(1 - p)}{p^2} + \frac{1}{p} - \frac{1}{p^2}$$

$$= \frac{2 - 2p + p - 1}{p^2}$$

$$= \frac{1 - p}{p^2}$$

5 a, d $f(x) = 0.8e^{-0.8x}, \quad x \geq 0$



b i $E(X) = \frac{1}{\lambda}$

$$= \frac{1}{0.8}$$

$$= 1.25$$

ii $\text{Var}(X) = \frac{1}{\lambda^2}$

$$\approx 1.56$$

c We need to find m such that

$$\int_0^m \lambda e^{-\lambda x} dx = \frac{1}{2}$$

$$\therefore \lambda \left[\frac{1}{-\lambda} e^{-\lambda x} \right]_0^m = \frac{1}{2}$$

$$\therefore -[e^{-\lambda m} - 1] = \frac{1}{2}$$

$$\therefore -e^{-\lambda m} + 1 = \frac{1}{2}$$

$$\therefore e^{-\lambda m} = \frac{1}{2}$$

$$\therefore e^{\lambda m} = 2$$

$$\therefore \lambda m = \ln 2$$

$$\therefore m = \frac{\ln 2}{\lambda}$$

d $m = \frac{\ln 2}{0.8} \approx 0.866$

e $F(X) = P(X \leq x)$

$$= \int_0^x \lambda e^{-\lambda t} dt$$

$$= \lambda \left[\frac{1}{-\lambda} e^{-\lambda t} \right]_0^x$$

$$= -(e^{-\lambda x} - e^0)$$

$$= 1 - e^{-\lambda x}$$

f $P(X > 1.3) = 1 - P(X \leq 1.3)$

$$= 1 - (1 - e^{-0.8 \times 1.3})$$

$$= e^{-1.04}$$

$$\approx 0.353$$

6 Let X be the volume of a bottle in mL. $X \sim N(376, 1.84^2)$

Then $\bar{X}$ is the average volume of each sample of 12.

$$\bar{X} \sim N\left(376, \frac{1.84^2}{12}\right)$$

- a $P(X < 373) \approx 0.0515$
 $\therefore$ about 5.15% will have a volume less than 373 mL.
- b $P(\bar{X} < 375) \approx 0.0299$
 $\therefore$ about 3% of all packs of 12 will have an average contents less than 375 mL.
- c From a and b there is a smaller chance of picking a 12-pack that does not meet the rules than that for an individual bottle. Hence, would prefer method II.
- d Let $X \sim N(\mu, 1.84^2)$

$$\text{We want } P(\bar{X} < 375) = 0.01$$

$$\therefore P\left(\frac{\bar{X} - \mu}{\frac{1.84}{\sqrt{12}}} < \frac{375 - \mu}{\frac{1.84}{\sqrt{12}}}\right) = 0.01$$

$$\therefore P\left(Z < \frac{375 - \mu}{\frac{1.84}{\sqrt{12}}}\right) = 0.01$$

$$\text{Thus } \frac{(375 - \mu)\sqrt{12}}{1.84} \approx -2.3263$$

$$\therefore 375 - \mu \approx -1.23567$$

$$\therefore \mu \approx 376.23 \dots$$

So, need to set it at $\mu = 377$ mL.

7 a As $\sigma^2 = 151.4 \text{ g}^2$, $\sigma \approx 12.304$ is known.
 The 95% CI for μ is:

$$\bar{x} - z_{0.025} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.025} \frac{\sigma}{\sqrt{n}}$$

Using technology, $594.5 \leq \mu \leq 598.9$ which means that we are 95% confident that the population lies within this region.

b As 600 does not lie in this interval, the sample data does not support the manufacturer's claim. It seems that the machine which fills the packets should be adjusted to add more contents to each packet.

c From the 95% CI in a,

$$-z_{0.025} \frac{\sigma}{\sqrt{n}} \leq \mu - \bar{x} \leq z_{0.025} \frac{\sigma}{\sqrt{n}}$$

$$\therefore |\mu - \bar{x}| \leq z_{0.025} \frac{\sigma}{\sqrt{n}}$$

$$\text{So we require } z_{0.025} \frac{\sigma}{\sqrt{n}} = 2$$

$$\text{Thus } 1.960 \times \frac{\sqrt{151.4}}{\sqrt{n}} = 2$$

$$\therefore \sqrt{\frac{151.4}{n}} = 1.0204$$

$$\begin{aligned}\therefore n &= \frac{151.4}{1.0204^2} \\ \therefore n &= 145.406 \dots \\ \therefore \text{a sample of 146 should be used.}\end{aligned}$$

8 (1) $H_0: \mu = \text{€}438\,000$

$H_1: \mu \neq \text{€}438\,000$

(2) The null distribution is $t(29)$ as $n = 30$.

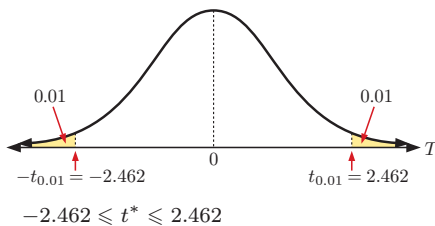
$$s_{n-1} = \sqrt{\frac{n}{n-1}} s_n = \sqrt{\frac{30}{29}} \times 23\,500 = \text{€}23\,902$$

is an unbiased estimate of σ .

(3) The test statistic is

$$t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} = \frac{(\bar{x} - 438\,000)\sqrt{30}}{23\,902}$$

(4) We reject H_0 if t^* lies in the critical region.



$$\therefore -2.462 \leq \frac{(\bar{x} - 438\,000)\sqrt{30}}{23\,902} \leq 2.462$$

$$\therefore -10\,744 \leq \bar{x} - 438\,000 \leq 10\,744$$

$$\therefore 427\,000 < \bar{x} < 448\,500$$

So, the real estate agent's claim would be supported if $\bar{x}$ lies between €427 000 and €448 500.

9 a $\text{Cov}(X, Y)$

$$\begin{aligned}&= E((X - \mu_X)(Y - \mu_Y)) \\ &= E(XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y) \\ &= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X \mu_Y \\ &= E(XY) - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

b If X and Y are independent then

$$\text{Cov}(X, Y) = 0$$

$$\therefore E(XY) - E(X)E(Y) = 0 \quad \{\text{from a}\}$$

$$\therefore E(XY) = E(X)E(Y)$$

c $\text{Cov}(X, X - Y)$

$$\begin{aligned}&= E(X(X - Y)) - E(X)E(X - Y) \\ &= E(X^2 - XY) - E(X)[E(X) - E(Y)] \\ &= E(X^2) - E(XY) - [E(X)]^2 + E(X)E(Y) \\ &= E(X^2) - [E(X)]^2 - E(X)E(Y) + E(X)E(Y) \\ &\quad \{E(XY) = E(X)E(Y) \text{ from b}\} \\ &= \text{Var}(X)\end{aligned}$$

10 a $X \sim N(\mu, 3.71^2)$

$$\therefore \bar{X} \sim N\left(\mu, \frac{3.71^2}{13}\right) \quad \{\text{Central Limit Theorem}\}$$

b If $\bar{X}$ lies in the critical region, we reject H_0 in favour of accepting H_1 .

c $\alpha = P(\text{Type I error})$

$$= P(\text{rejecting } H_0 \mid H_0 \text{ is true})$$

$$= P\left(\bar{X} < 498 \text{ or } \bar{X} > 502 \mid \bar{X} \sim N\left(500, \frac{3.71^2}{13}\right)\right)$$

$$= 1 - P\left(498 \leq \bar{X} \leq 502 \mid \bar{X} \sim N\left(500, \frac{3.71^2}{13}\right)\right)$$

$$\approx 0.0519$$

d $P(\text{Type II error})$

$$= P(\text{accepting } H_0 \mid H_1 \text{ is true})$$

$$= P\left(498 \leq \bar{X} \leq 502 \mid \bar{X} \sim N\left(498.4, \frac{3.71^2}{13}\right)\right)$$

$$= 0.651$$

11 (1) $H_0: \mu = 106.3$

$H_1: \mu > 106.3$

(2) As $\sigma = 12.41$, the null distribution is Z .

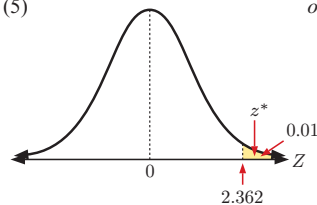
(3) The test statistic is z^* where

$$z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{110.1 - 106.3}{\frac{12.41}{\sqrt{65}}} \approx 2.469$$

(4) We reject H_0 if:

z^* lies in the critical region or the p -value is < 0.01

(5)



or p -value
= 0.00678

which is < 0.01

(6) We reject H_0 .

or We reject H_0 .

(7) There is sufficient evidence at a 1% level of significance to suggest that the mean weight of a tomato has increased.

12 a $E(T_1) = \frac{3}{10}E(X_1) + \frac{7}{10}E(X_2)$

$$= \frac{3}{10}\mu + \frac{7}{10}\mu$$

$$= \mu$$

$\therefore T_1$ is an unbiased estimator of μ .

b $E(T_2) = \frac{3}{5}E(X_1) + \frac{2}{5}E(X_2)$

$$= \frac{3}{5}\mu + \frac{2}{5}\mu$$

$$= \mu$$

$\therefore T_2$ is an unbiased estimator of μ .

c i Lucy's estimate = $\frac{3(2.6) + 7(5.3)}{10}$

$$= 4.49$$

ii Beth's estimate = $\frac{3(2.6) + 2(5.3)}{5}$

$$= 3.68$$

iii $\text{Var}(T_1) = \frac{9}{100}\text{Var}(X_1) + \frac{49}{100}\text{Var}(X_2)$

$$= \frac{9}{100}\sigma^2 + \frac{49}{100}\sigma^2$$

$$= \frac{58}{100}\sigma^2$$

$$\begin{aligned}\text{Var}(T_2) &= \frac{9}{25} \text{Var}(X_1) + \frac{4}{25} \text{Var}(X_2) \\ &= \frac{9}{25} \sigma^2 + \frac{4}{25} \sigma^2 \\ &= \frac{13}{25} \sigma^2 \\ &= \frac{52}{100} \sigma^2 \quad \text{which is } < \text{Var}(T_1)\end{aligned}$$

$\therefore$ Eve is not correct.

T_2 is a more efficient estimator than T_1 .

$$\begin{aligned}\text{d } E(T_3) &= \frac{a}{c} E(X_1) + \frac{b}{c} E(X_2) \\ &= \frac{a}{c} \mu + \frac{b}{c} \mu \\ &= \left(\frac{a+b}{c} \right) \mu \\ &= \mu \\ \therefore a+b &= c\end{aligned}$$

REVIEW SET C

1 a $Y = 2X_3 - 2X_2 - X_1$

$$\begin{aligned}\text{i } E(Y) &= 2E(X_3) - 2E(X_2) - E(X_1) \\ &= 2a - 2(3) - (2) \\ &= 2a - 8\end{aligned}$$

$$\begin{aligned}\text{ii } \text{Var}(Y) &= 2^2 \text{Var}(X_3) + (-2)^2 \text{Var}(X_2) + (-1)^2 \text{Var}(X_1) \\ &= 4b + 4 \left(\frac{1}{16} \right) + \frac{1}{8} \\ &= 4b + \frac{1}{4} + \frac{1}{8}\end{aligned}$$

b If $E(Y) = 0$, $2a - 8 = 0$
 $\therefore a = 4$

If $\text{Var}(Y) = 1$, $4b + \frac{3}{8} = 1$
 $\therefore 4b = \frac{5}{8}$
 $\therefore b = \frac{5}{32}$

c Y is a linear combination of normal variables. Consequently Y is Normal.
So, $Y = N(0, 1)$.

d $P(X_3 \geq 8b) = P(X_3 \geq 1.25)$
 ≈ 0.106

2 $P(\text{'six'}) = \frac{1}{6}$

Let X be the number of rolls needed to obtain a 'six'.

$$\therefore X \sim \text{Geo}\left(\frac{1}{6}\right) \quad \text{and} \quad \mu_X = \frac{1}{p} = \frac{1}{\frac{1}{6}} = 6$$

So, on average a player takes 6 rolls to win €10.

As Pierre wants to make €2/game, he must charge €12 over the 6 rolls, which is €2 per roll.

3 a $X \sim \text{Po}(m)$ where m is the unknown mean number of errors per page.

$$\begin{aligned}\text{b } P(X = x) &= \frac{m^x e^{-m}}{x!} \quad \text{for } x = 0, 1, 2, 3, \dots \\ \text{i } P(X = 1) &= \frac{m e^{-m}}{1!} = q(-\ln q) = -q \ln q \\ \text{ii } P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - q + q \ln q\end{aligned}$$

c i Let Y denote returns based on numbers of errors.

y	\$10	\$1	-\$8
$P(Y = y)$	q	$-q \ln q$	$1 - q + q \ln q$

$$\begin{aligned}\text{ii } E(Y) &= \sum y_i p_i \\ &= \$[10q - q \ln q - 8(1 - q + q \ln q)] \\ &= \$[18q - 9q \ln q - 8]\end{aligned}$$

iii To receive a bonus (positive) we require $18q - 9q \ln q - 8 > 0$.

$$\begin{aligned}\text{First we solve } 18q - 9q \ln q - 8 &= 0 \\ \therefore q &\approx 0.268\end{aligned}$$

$$\text{For } q = 0.267, \quad f(q) \approx -0.0208 < 0$$

$$\begin{aligned}\text{For } q = 0.268, \quad f(q) &\approx 0.000\,045\,1 \\ \text{which is } &> 0\end{aligned}$$

$$\therefore \text{required } q = 0.268$$

4 a We know that $\int_0^1 f(x) dx = 1 \dots (1)$

for a well-defined pdf, and that

$$E(X) = \int_0^1 x f(x) dx = 0.7 \dots (2)$$

$$\text{From (1), } \int_0^1 (ax^3 + bx^2) dx = 1$$

$$\begin{aligned}\therefore \left[\frac{ax^4}{4} + \frac{bx^3}{3} \right]_0^1 &= 1 \\ \therefore \frac{a}{4} + \frac{b}{3} - 0 &= 1 \\ \therefore 3a + 4b &= 12 \dots (3)\end{aligned}$$

$$\text{From (2), } \int_0^1 (ax^4 + bx^3) dx = 0.7$$

$$\begin{aligned}\therefore \left[\frac{ax^5}{5} + \frac{bx^4}{4} \right]_0^1 &= \frac{7}{10} \\ \therefore \frac{a}{5} + \frac{b}{4} - 0 &= \frac{7}{10} \\ \therefore 4a + 5b &= 14 \dots (4)\end{aligned}$$

Solving (3) and (4) simultaneously gives $a = -4$, $b = 6$.

b $P(\text{runs out of petrol}) = P(X > 0.95)$

$$\begin{aligned}&= \int_{0.95}^1 (-4x^3 + 6x^2) dx \\ &\approx 0.0998\end{aligned}$$

$\therefore$ the service station runs out of petrol about 10% of the time.

5 a As $\mu = \frac{pr}{1-p}$, $\mu - \mu p = pr$

$$\therefore \mu = pr + \mu p$$

$$\therefore \mu = p(\mu + r)$$

$$\therefore p = \frac{\mu}{\mu + r}$$

- b** Hence, $G(t) = \left(\frac{1 - \frac{\mu}{\mu+r}}{1 - \frac{\mu t}{\mu+r}} \right)^r$
- $$= \left(\frac{\mu + r - \mu}{\mu + r - \mu t} \right)^r$$
- $$= \left(\frac{r}{\mu + r - \mu t} \right)^r$$
- $$= \left(\frac{1}{1 + \frac{\mu}{r}(1-t)} \right)^r$$
- $$= \frac{1}{\left(1 + \frac{\mu(1-t)}{r} \right)^r}$$
- c** As $r \rightarrow \infty$, $\left(1 + \frac{\mu(1-t)}{r} \right)^r \rightarrow e^{\mu(1-t)}$
- $$\left\{ \lim_{n \rightarrow \infty} \left(1 + \frac{a}{n} \right)^n \rightarrow e^a, \quad \mu \text{ and } t \text{ fixed} \right\}$$
- $$\therefore \lim_{r \rightarrow \infty} G(t) = e^{-\mu(1-t)}$$
- $$= e^{\mu(t-1)}$$
- which is the PGF for the Poisson variable with mean μ .
- d** $\text{Po}(m) = \lim_{r \rightarrow \infty} \text{NB} \left(r, \frac{m}{m+r} \right)$

- 6 a** $X \sim N(63.4, 40.1)$
 $P(X \leq 60) \approx 0.296$
 About 29.6% of sausages produced weigh ≤ 60 grams.
- b** $\bar{W} \sim N(63.4, \frac{40.1}{10})$ {CL Theorem}
- c** $P(\bar{W} \leq 60) \approx 0.0448$
 In approximately 4.48% of all samples of 10 sausages, the mean weight in the sample will be ≤ 60 g.
- d** We require $P(\bar{W} \leq 60) = 0.01$
- $$P \left(\frac{\bar{W} - 63.4}{\sqrt{\frac{40.1}{n}}} \leq \frac{60 - 63.4}{\sqrt{\frac{40.1}{n}}} \right) = 0.01$$
- $$\therefore P \left(Z \leq \frac{-3.4}{\sqrt{\frac{40.1}{n}}} \right) = 0.01$$
- $$\therefore \frac{-3.4\sqrt{n}}{\sqrt{40.1}} \approx -2.3263$$
- $$\therefore \sqrt{n} \approx 4.333$$
- $$\therefore n \approx 18.77$$
- $\therefore$ samples of size 19 are needed.

- 7 a** As σ is unknown, the t -distribution replaces the Z -distribution for the CI and s_{n-1} is used as an unbiased estimate of σ .
 Thus, the 95% CI for μ is:

$$\bar{x} - t(0.025, n-1) \frac{s_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{x} + t(0.025, n-1) \frac{s_{n-1}}{\sqrt{n}}$$

- b** Adding the two outer boundary limits gives $2\bar{x}$.

$$\therefore 2\bar{x} = 139.91 + 147.49 = 287.4$$

$$\therefore \bar{x} = 143.7$$

- c** In order to find n we would need to solve

$$\bar{x} + t(0.025, n-1) \frac{s_{n-1}}{\sqrt{n}} = 147.49$$

$$\therefore \frac{t(0.025, n-1)s_{n-1}}{\sqrt{n}} = 3.79$$

$$\therefore \frac{t(0.025, n-1)\sqrt{\frac{s_{n-1}^2}{n-1}}}{\sqrt{n}} = 3.79$$

$$\therefore \frac{t(0.025, n-1) \times 11.2}{\sqrt{n-1}} = 3.79$$

$$\therefore t(0.025, n-1) = 0.3384\sqrt{n-1}$$

which is not solvable algebraically as n must be known for a t calculation.

$$\mathbf{d} \quad n-1 = \left[\frac{t(0.025, n-1)}{0.3384} \right]^2$$

$$\therefore n = \left[\frac{t(0.025, n-1)}{0.3384} \right]^2 + 1$$

$$\text{If } n = 33, \quad n \approx 37.23 \quad \times$$

$$\text{If } n = 34, \quad n \approx 37.14 \quad \times$$

$$\text{If } n = 35, \quad n \approx 35.06 \quad \checkmark$$

$$\text{If } n = 36, \quad n \approx 36.90 \quad \times$$

Thus $n = 35$.

- 8** Let X_1 be the number of fish caught before the course, and X_2 be the number of fish caught after the course.

We consider $D = X_2 - X_1$

a $H_0: \mu_D = 0$ against

$H_1: \mu_D > 0$ (the course was effective)

D values are: 12, 9, 18, -3, -9, 4, 0, 10, 4

where $\bar{d} = 5$ and $s_{n-1} \approx 8.2614$

As σ is unknown, we use s_{n-1} as an unbiased estimate of σ .

$$\text{The test statistic is } t^* = \frac{\bar{d} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{5 - 0}{\frac{8.2614}{\sqrt{9}}}$$

$$\therefore t^* \approx 1.816$$

$$\text{and } p\text{-value} \approx 0.0535$$

which is > 0.05

The decision:

- as $p\text{-value} > 0.05$ or
- as t^* does not lie in the rejection region ($t > 1.860$) then we do not reject H_0 and are subject to making a Type II error, that is, accepting H_0 when it is in fact false.

Note: We do not have enough information to determine the probability of making this type of error.

- b** A 90% confidence interval for the mean difference is $]-0.121, 10.121[$ and as the null hypothesis value of $\mu_D = 0$ is within the CI, then at a 5% level, this is consistent with the acceptance of H_0 .

9 (1) $H_0: \mu = 200$ and $H_1: \mu < 200$

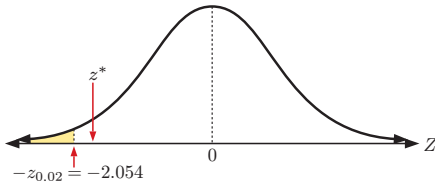
(2) The null distribution is Z .

(3) The test statistic is z^* where

$$z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{196.4 - 200}{\frac{11.36}{\sqrt{35}}} \\ \therefore z^* \approx -1.875$$

(4) We reject H_0 if z^* lies in the critical region.

(5)



(6) Since z^* does not lie in the critical region we do not reject H_0 .

(7) There is insufficient evidence to reject H_0 , that is, there is insufficient evidence to suggest that her present mean has fallen to less than 200.

10 a X is a discrete random variable with values 0, 1, 2, 3, 4,
 $\therefore S$ is also a discrete random variable and the critical region is $\{0 \leq S \leq 33\} \cup \{S \geq 57\}$.

b A Type I error is rejecting H_0 when H_0 is in fact true.
 This means, deciding that $m \neq 3$ when in fact $m = 3$ (and $X \sim \text{Po}(3)$).

$$\begin{aligned} & \text{P(Type I error)} \\ &= \text{P}(S \text{ is in the critical region} \mid S \sim \text{Po}(15 \times 3)) \\ &= \text{P}(0 \leq S \leq 33 \text{ or } S \geq 57 \mid S \sim \text{Po}(45)) \\ &= \text{P}(S \leq 33 \mid S \sim \text{Po}(45)) \\ &\quad + 1 - \text{P}(S \leq 56 \mid S \sim \text{Po}(45)) \\ &\approx 0.038\,339 + 1 - 0.952\,745 \\ &\approx 0.0856 \end{aligned}$$

c A Type II error is accepting H_0 when H_0 is in fact false.
 This means accepting $m = 3$ when in fact $m = 3.4$.

$$\begin{aligned} & \text{P(Type II error)} \\ &= \text{P}(34 \leq S \leq 56 \mid S \sim \text{Po}(15 \times 3.4)) \\ &= \text{P}(S \leq 56 \mid S \sim \text{Po}(51)) - \text{P}(S \leq 33 \mid S \sim \text{Po}(51)) \\ &\approx 0.782\,27 - 0.004\,79 \\ &\approx 0.777 \end{aligned}$$

11 $H_0: \rho = 0$ (X and Y are independent) against

$H_1: \rho \neq 0$

$n = 27$ and so $\nu = 27 - 2 = 25$

We require the p -value to be > 0.05 for H_0 acceptance.

$$\therefore 2 \times \text{P}(T > |t^*|) > 0.05$$

$$\therefore \text{P}(T > |t^*|) > 0.025$$

$$\therefore \text{P}(T \leq |t^*|) < 0.975$$

$$\therefore t^* \approx 2.060 \quad \{\text{technology}\}$$

$$\text{But } t^* = r \sqrt{\frac{n-2}{1-r^2}}$$

$$\therefore \text{ we accept } H_0 \text{ when } r \sqrt{\frac{25}{1-r^2}} < 2.06$$

$$\therefore \frac{25r^2}{1-r^2} < 4.2436$$

Since $1 - r^2 > 0$,

$$\therefore 25r^2 < 4.2436 - 4.2436r^2$$

$$\therefore 29.2436r^2 < 4.2436$$

$$\therefore r^2 < 0.145\,11$$

$$\therefore \sqrt{r^2} < 0.380\,93 \dots$$

$$\therefore |r| < 0.380\,93 \dots$$

Thus for X, Y independence, the greatest value of $|r|$ is 0.3809 (to 4 s.f.).

12 a Using $S_{n-1}^2 = \frac{n}{n-1} S_n^2$

Sample A gives

$$S_{n-1}^2 = \frac{8}{7} \times 4.2 = 4.8$$

Sample B gives

$$S_{n-1}^2 = \frac{22}{21} \times 5.1 \approx 5.34$$

So, 4.8 and 5.34 are unbiased estimates of σ^2 .

b i $t_1 = \frac{8(4.2) + 22(5.1)}{30} = 4.86$

ii $E(T_1) = E\left(\frac{8}{30} S_A^2 + \frac{22}{30} S_B^2\right)$
 $= \frac{8}{30} E(S_A^2) + \frac{22}{30} E(S_B^2)$
 $= \frac{8}{30} \times \frac{7}{8} \sigma^2 + \frac{22}{30} \times \frac{21}{22} \sigma^2$
 $= \frac{28}{30} \sigma^2$
 $\neq \sigma^2$

$\therefore T_1$ is a biased estimator of σ^2 , and so t_1 is a biased estimate of σ^2 .

c $E(T_2) = \frac{a}{c} E(S_A^2) + \frac{b}{c} E(S_B^2)$
 $= \frac{a}{c} \times \frac{7}{8} \sigma^2 + \frac{b}{c} \times \frac{21}{22} \sigma^2$
 $= \left(\frac{7a}{8c} + \frac{21b}{22c}\right) \sigma^2$

which is σ^2 if $\frac{7a}{8c} + \frac{21b}{22c} = 1$
 $\therefore 77a + 84b = 88c$

REVIEW SET D

1 Let S be the volume of a small bottle and let
 L be the volume of a large bottle.

$$\therefore S \sim N(338, 3^2) \text{ and } L \sim N(1010, 12^2)$$

a Consider $U = L - (S_1 + S_2 + S_3)$
 $= L - S_1 - S_2 - S_3$

$$\begin{aligned} \therefore E(U) &= E(L) - E(S_1) - E(S_2) - E(S_3) \\ &= 1010 - 3 \times 338 \\ &= -4 \text{ mL} \end{aligned}$$

$$\begin{aligned} \text{and } \text{Var}(U) &= \text{Var}(L) + \text{Var}(S_1) + \text{Var}(S_2) + \text{Var}(S_3) \\ &= 12^2 + 3 \times 3^2 \\ &= 171 \text{ mL}^2 \end{aligned}$$

$$\begin{aligned} & \text{P}(L > S_1 + S_2 + S_3) \\ &= \text{P}(L - S_1 - S_2 - S_3 > 0) \\ &= \text{P}(U > 0) \text{ where } U \sim N(-4, 171) \\ &\approx 0.380 \end{aligned}$$

b Consider $V = L - 3S$

$$\begin{aligned}\therefore E(V) &= E(L) - 3E(S) \\ &= 1010 - 3 \times 338 \\ &= -4\end{aligned}$$

$$\begin{aligned}\text{and } \text{Var}(V) &= \text{Var}(L) + 9 \text{Var}(S) \\ &= 12^2 + 9 \times 3^2 \\ &= 225\end{aligned}$$

$$\therefore V \sim N(-4, 225)$$

$$\begin{aligned}P(L > 3S) &= P(L - 3S > 0) \\ &= P(V > 0) \\ &\approx 0.395\end{aligned}$$

2 a Let X be the number of patients arriving between 9:00 am and 9:45 am.

$$E(X) = \frac{3}{4} \text{ of } 14 = 10.5$$

$$X \sim \text{Po}(10.5)$$

$$\therefore P(X = 5) = \frac{10.5^5 e^{-10.5}}{5!} \approx 0.0293$$

b Let Y be the number of patients arriving between 10:00 am and 10:30 am.

$$E(Y) = \frac{1}{2} \text{ of } 14 = 7$$

$$Y \sim \text{Po}(7)$$

$$\begin{aligned}\therefore P(Y < 7) &= P(Y \leq 6) \\ &\approx 0.450\end{aligned}$$

3 a As $E(X) = np = 8$ and

$$\text{Var}(X) = np(1-p) = 6,$$

$$8(1-p) = 6$$

$$\therefore 1-p = \frac{3}{4}$$

$$\therefore p = \frac{1}{4}$$

$$\text{Thus } n\left(\frac{1}{4}\right) = 8 \quad \therefore n = 32$$

b i $X \sim B(32, \frac{1}{4})$

$$\begin{aligned}\therefore P(X \geq 4) &= 1 - P(X \leq 3) \\ &\approx 0.975\end{aligned}$$

ii As $n \geq 30$, we can use the Central Limit Theorem.

$$\bar{X} \sim N(8, \frac{6}{32})$$

$$\therefore P(7.9 \leq \bar{X} \leq 8.1) \approx 0.406$$

4 As $f(x)$ is the normal PDF,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

$$\therefore \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \sqrt{2\pi}\sigma \quad \{\sigma \text{ is fixed}\}$$

Letting $\mu = 9$ and $\sigma = 2$ we get

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-9}{2}\right)^2} dx = 2\sqrt{2\pi}$$

$$\therefore \int_{-\infty}^{\infty} e^{-\frac{1}{8}(x-9)^2} dx = 2\sqrt{2\pi}$$

5 $G(t) = (1 - \beta \ln t)^{-\alpha-1}$, α and β are parameters

$$\begin{aligned}\text{a } G'(t) &= (-\alpha-1)(1 - \beta \ln t)^{-\alpha-2} \times -\beta \left(\frac{1}{t}\right) \\ &= (\alpha+1)\beta \left[\frac{(1 - \beta \ln t)^{-\alpha-2}}{t} \right]\end{aligned}$$

When $t = 1$, $\ln t = 0$

$$\therefore \mu = G'(1) = (\alpha+1)\beta \left[\frac{1}{1} \right] = (\alpha+1)\beta$$

b $G''(t)$

$$= (\alpha+1)\beta \left[\frac{(-\alpha-2)(1-\beta \ln t)^{-\alpha-3} \times \frac{-\beta}{t} - (1-\beta \ln t)^{-\alpha-2} \times 1}{t^2} \right]$$

{quotient rule}

$$\therefore G''(1) = (\alpha+1)\beta \left[\frac{\beta(\alpha+2) - 1}{1} \right]$$

$$\begin{aligned}\text{and } \text{Var}(X) &= G''(1) + G'(1) - [G'(1)]^2 \\ &= (\alpha+1)\beta^2(\alpha+2) - (\alpha+1)\beta \\ &\quad + (\alpha+1)\beta - (\alpha+1)^2\beta^2 \\ &= (\alpha+1)\beta^2[\alpha+2 - \alpha - 1] \\ &= (\alpha+1)\beta^2 \times 1 \\ &= (\alpha+1)\beta^2\end{aligned}$$

$$\text{6 a } s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{230}{15} \approx 15.33$$

b $s_{n-1} = \sqrt{\frac{n}{n-1}} s_n \approx 4.0532$ is an unbiased estimate of σ .

c i A CI for μ is $124.94 < \mu < 129.05$ and $\bar{x}$ for this sample is the midpoint of the CI.

$$\therefore \bar{x} = \frac{124.94 + 129.05}{2} = 126.995$$

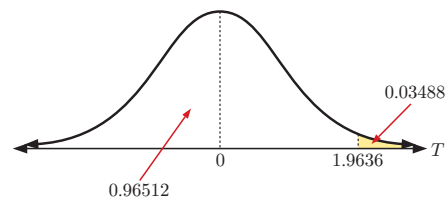
As σ is unknown (had to be estimated), we have a t -distribution with $\nu = 15 - 1 = 14$, that is, $t \sim t(14)$.

A 95% CI is $124.75 < \mu < 129.24$

$$\text{ii } t^* = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}} \approx \frac{|126.995 - 129.05|}{\frac{4.0532}{\sqrt{15}}}$$

$$\therefore t^* \approx 1.9636 \dots$$

$$\text{and } P(T < 1.9636 \dots) \approx 0.96512$$



$$\therefore \alpha \approx 2 \times 0.3488$$

$$\therefore \alpha \approx 6.98$$

$\therefore$ we have a 7% confidence level.

7 a B, as:

- A is closer to normality
- A's standard deviation is much less than 9.21

b $\mu_{\bar{X}} = \mu = 11.4$

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{9.21^2}{9} \approx 9.425$$

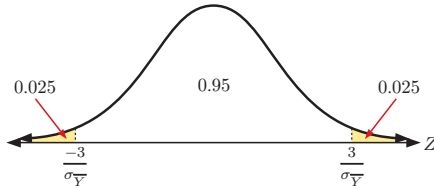
c $P(\bar{X} \geq 12) \approx 0.423$

d Firstly we need to consider

$$P(\mu_{\bar{Y}} - 3 \leq \bar{Y} \leq \mu_{\bar{Y}} + 3) = 0.95$$

$$\therefore P\left(\frac{\mu_{\bar{Y}} - 3 - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} \leq \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} \leq \frac{\mu_{\bar{Y}} + 3 - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}}\right) = 0.95$$

$$\therefore P\left(\frac{-3}{\sigma_{\bar{Y}}} \leq Z \leq \frac{3}{\sigma_{\bar{Y}}}\right) = 0.95$$



$$\therefore P\left(Z \leq -\frac{3}{\sigma_{\bar{Y}}}\right) = 0.025$$

$$\therefore -\frac{3}{\sigma_{\bar{Y}}} \approx -1.960$$

$$\therefore \sigma_{\bar{Y}} \approx 1.531$$

$$\text{But } \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$$

$$\therefore n = \frac{\sigma_Y^2}{\sigma_{\bar{Y}}^2}$$

$$\therefore n \approx 36.2$$

So, n must be at least 37.

8 (1) $H_0: \mu = 546$ (there is no change)

$H_1: \mu > 546$ (there is an increase)

(2) As σ^2 is unknown we use the t -distribution with $\nu = 49$.

(3) The test statistic is t^* where

$$t^* = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} = \frac{563 - 546}{\frac{59.049}{\sqrt{50}}}$$

$$\begin{aligned} \{s_{n-1} &= \sqrt{\frac{n}{n-1}} s_n \\ &= \sqrt{\frac{50}{49}} \times \sqrt{3417} \\ &\approx 59.049\} \end{aligned}$$

$$\therefore t^* \approx 2.0357$$

(4) We reject H_0 if:

t^* is in the critical region

or the p -value is < 0.05

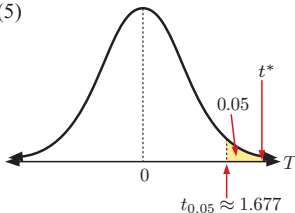
(5)

or

p -value

$$= P(T \geq 2.0357)$$

$$\approx 0.0236$$



(6) As t^* lies in the critical region or As the p -value is < 0.05 , we reject H_0 .

(7) There is sufficient evidence to reject H_0 in favour of H_1 . We accept that the new brand has longer life at a 5% significance level.

9 a As $X \sim \text{Geo}(0.25)$, $S \sim \text{NB}(12, 0.25)$

b H_0 should be accepted if $32 \leq S \leq 74$ $\{S \geq 12\}$.

c $\alpha = P(\text{Type I error})$

$= P(\text{rejecting } H_0 \mid H_0 \text{ is in fact true})$

$= P(S \leq 31 \text{ or } S \geq 75 \mid S \sim \text{NB}(12, 0.25))$

$= P(S \leq 31 \mid S \sim \text{NB}(12, 0.25))$

$+ 1 - P(S \leq 74 \mid S \sim \text{NB}(12, 0.25))$

$$= \sum_{i=12}^{31} \binom{i-1}{11} (0.25)^{12} (0.75)^{i-12}$$

$$+ 1 - \sum_{i=12}^{74} \binom{i-1}{11} (0.25)^{12} (0.75)^{i-12}$$

$$\approx 0.064426 + 1 - 0.974523$$

$$\approx 0.08990$$

Thus $\alpha \approx 0.9$

d $P(\text{Type II error})$

$= P(\text{accepting } H_0 \mid H_1 \text{ is true})$

$= P(32 \leq S \leq 74 \mid S \sim \text{NB}(12, 0.2))$

$= P(S \leq 74 \mid S \sim \text{NB}(12, 0.2))$

$- P(S \leq 31 \mid S \sim \text{NB}(12, 0.2))$

$$= \sum_{i=12}^{74} \binom{i-1}{11} (0.2)^{12} (0.8)^{i-12}$$

$$- \sum_{i=12}^{31} \binom{i-1}{11} (0.2)^{12} (0.8)^{i-12}$$

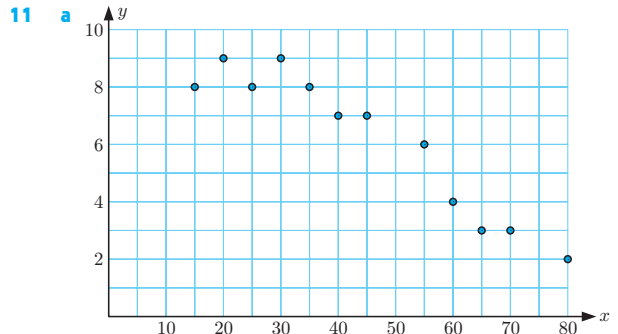
$$\approx 0.83084 - 0.01272$$

$$\approx 0.818$$

$$\therefore \text{power} \approx 1 - 0.818 \approx 0.182 \text{ or } 18.2\%$$

10 a A Type I error would result if it was determined that Quickchick is supplying underweight chickens when they are in fact not.

b A Type II error would result if Quickchick is supplying underweight chickens when it is determined that they are not.



b $r \approx -0.9462$ {technology}

c The correlation is strong and negative.

d Yes, as the distance from the target does affect the number of hits.

e $y = -0.1134x + 11.27$ hits

f When $x = 50$, $y = -0.1134 \times 50 + 11.27$
 $= 5.6$

So, on average we expect about 5.6 hits from a 50 m distance, for every 10 arrows shot.

$\therefore$ for 100 arrows we predict $5.6 \times 10 = 56$ hits.

g No, as 100 m is outside the data set and so any calculation would be unreliable.

h $x \approx 7.894y + 93.68$ metres

12 a $E(T_1) = \frac{2}{3}E(X_1) + \frac{1}{3}E(X_2)$

$$= \frac{2}{3}\mu + \frac{1}{3}\mu$$

$$= \mu$$

$$E(T_2) = aE(X_1) + (1-a)E(X_2)$$

$$= a\mu + (1-a)\mu$$

$$= a\mu + \mu - a\mu$$

$$= \mu$$

$\therefore T_1$ and T_2 are both unbiased estimators of μ .

b $\text{Var}(T_1) = \frac{4}{9}\text{Var}(X_1) + \frac{1}{9}\text{Var}(X_2)$

$$= \frac{4}{9}\sigma^2 + \frac{1}{9}\sigma^2$$

$$= \frac{5}{9}\sigma^2$$

$$\text{Var}(T_2) = a^2\text{Var}(X_1) + (1-a)^2\text{Var}(X_2)$$

$$= a^2\sigma^2 + (1-a)^2\sigma^2$$

$$= (2a^2 - 2a + 1)\sigma^2$$

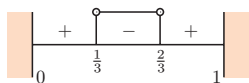
c If T_2 is a more efficient estimator than T_1

$$\text{Var}(T_2) < \text{Var}(T_1)$$

$$\therefore 2a^2 - 2a + 1 < \frac{5}{9}$$

$$\therefore a^2 - a + \frac{2}{9} < 0$$

$$\therefore (a - \frac{2}{3})(a - \frac{1}{3}) < 0$$

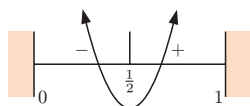


$$\therefore \frac{1}{3} < a < \frac{2}{3} \quad \text{or} \quad a \in]\frac{1}{3}, \frac{2}{3}[$$

d $\text{Var}(T_2)$ is a minimum when $2a^2 - 2a + 1 = A$ is a minimum.

$$\frac{dA}{da} = 4a - 2 \quad \text{which is 0 when } a = \frac{1}{2}$$

Sign diagram:



$\therefore$ a minimum when $a = \frac{1}{2}$.

13 a $E(U) = 3E(X) - 5E(Y)$

$$= 3\mu_X - 5\mu_Y$$

and $\text{Var}(U) = 9\text{Var}(X) + 25\text{Var}(Y)$

$$= 9\sigma_X^2 + 25\sigma_Y^2$$

b $E(aS_X^2 + bS_Y^2)$

$$= aE(S_X^2) + bE(S_Y^2)$$

$$= a\sigma_X^2 + b\sigma_Y^2$$

$$= 9\sigma_X^2 + 25\sigma_Y^2$$

$\therefore a = 9$ and $b = 25$

INDEX

- | | | | |
|--|----------------------------|--|----------------|
| alternative hypothesis | 100 | mean line | 126, 131 |
| Bernoulli distribution | 26, 37, 56 | mean point | 126, 131 |
| Bernoulli random variable | 26 | method of least squares | 135 |
| bias | 68 | modal value | 30 |
| biased estimate | 83, 86 | negative binomial distribution | 31, 37, 56, 60 |
| biased estimator | 83, 86 | negative binomial random variable | 31 |
| binomial distribution | 27, 37, 48, 56, 60 | negative linear correlation | 125, 128 |
| binomial formula | 52 | normal distribution | 20, 47 |
| binomial random variable | 27 | normal random variable | 20, 47 |
| binomial series | 52 | null distribution | 102 |
| bivariate normal distribution | 139 | null hypothesis | 100 |
| bivariate statistics | 124 | one-tailed test | 103 |
| causation | 124 | paired data | 98 |
| Central Limit Theorem | 71, 90 | parameter | 66 |
| confidence interval | 66, 90 | perfect linear correlation | 125, 133 |
| confidence level | 90, 92, 93 | Poisson distribution | 33, 37, 56, 60 |
| continuous exponential random variable | 44, 47 | Poisson random variable | 33 |
| continuous normal random variable | 20, 47 | population proportion | 79 |
| continuous random variable | 9, 42 | positive linear correlation | 125, 128 |
| continuous uniform distribution | 43, 47 | power of a test | 115 |
| continuous uniform random variable | 43, 47 | probability density function | 9 |
| correction for continuity | 48 | probability distribution function | 9, 42 |
| correlation | 124 | probability generating function | 52 |
| covariance | 131, 137 | probability mass function | 9 |
| critical region | 103, 108 | product moment correlation coefficient | 126, 132 |
| critical value | 103, 108 | proportion | 79 |
| cumulative distribution function | 9, 42 | p-value | 104 |
| degrees of freedom | 86, 94, 142 | random error | 68 |
| discrete random variable | 9, 26 | random sampling | 66 |
| discrete uniform distribution | 26, 37, 56, 60 | random variable | 9 |
| discrete uniform random variable | 26 | regression coefficient | 135 |
| estimate | 82 | regression line | 135 |
| estimator | 82 | sample error | 73 |
| expectation | 10 | sample mean | 70, 85 |
| expected value | 10 | sample proportion | 79 |
| exponential distribution | 44, 47 | sample size | 73 |
| exponential random variable | 44, 47 | sample variance | 85 |
| exponential series | 52 | sampling distribution | 66, 70, 83 |
| finite geometric series | 52 | sampling error | 73 |
| geometric discrete random variable | 29 | scatter diagram | 124 |
| geometric distribution | 29, 37, 56, 60 | standard deviation | 66 |
| hypotheses | 100 | standard error | 73 |
| hypothesis tests | 100, 106, 142 | standard normal distribution | 14 |
| independent random variables | 17, 20, 61, 62, 132, 133 | standard normal random variable | 14, 47 |
| infinite geometric series | 52 | standardised variable | 14 |
| joint probability distribution | 124, 139 | statistic | 66 |
| level of significance | 103 | statistical error | 68 |
| line of regression | 135 | statistical hypothesis | 100 |
| linear correlation | 125 | Student's t-distribution | 93, 142 |
| linear transformation | 14 | systematic error | 68 |
| matched pairs | 98, 112 | t-distribution | 93 |
| mean | 10, 36, 37, 42, 58, 60, 66 | test statistic | 102, 108 |
| | | two-tailed test | 102, 108 |
| | | Type I error | 101, 114, 118 |
| | | Type II error | 101, 114, 118 |

unbiased estimate	83, 86
unbiased estimator	83, 86
uncorrelated	125, 127
uniform distribution (continuous)	43, 47
uniform distribution (discrete)	26, 37, 56, 60
variance	13, 36, 37, 42, 58, 60

